Al-Mustansiriyah Journal of Science
ISSN: 1814-635X (print), ISSN:2521-3520 (online) Volume 33, Issue 3, 2022 DOI: http://doi.org/10.23851/mjs.v33i3.1069

Research Article **Open Access**

Information Retrieval for Cancer Cell Detection Based on Advanced Machine Learning Techniques

Atheel Sabih Shaker[1], Saadaldeen Rashid Ahmed[2*]

[1] Computer Engineering Techniques, Baghdad College of Economic Sciences University, Baghdad, IRAQ.
[2] Computer Science, College of Computer Science and Math, University of Tikrit, IRAQ.

*Correspondent contact: Saadaljanabi95@gmail.com

Article Info

Received
07/08/2021

Accepted
24/04/2022

Published
25/09/2022

ABSTRACT

In this research paper, we focus on designing and developing a fully automated gene regulation from cancerous cell heterogeneity using advanced machine learning techniques. There are several modern technologies developed to make DNA sequencing easier and cheaper. Among them, gene regulation produces the longest read sequences and the lengths of the reads are growing day by day. Machine learning technique like Support Vector Machine (SVM) is developed to align these gene sequences. Every technique faced some challenges, but facing the desired challenges introduce some new challenges on the other side. So, no one tool is perfect for every work. The SVM technique is a new aligner tool that does a tradeoff and performs better from different aspects. For the model with the best generator loss, an average maximum validation accuracy of 91.29% was achieved. The gene regulation with SVM is like a mini-map that takes a few times more memory to index the whole genome of a reference sequence. The single-cell data are the main target of SVM. It is shown that it would help the SVM and similar techniques to align better with long insertions and deletions of single-cell gene regulation. Single-cell data is run against the well-known reference sequence and a randomly generated synthetic reference.

KEYWORDS: Information; Retrieval; Single Cell; Support Vector Machine; Machine Learning, RNA; DNA.

الخلاصة

في هذه الورقة البحثية، نركز على تصميم وتطوير تنظيم جيني مؤتمت بالكامل من بيانات الخلية باستخدام تقنيات التعلم الآلي المتقدمة. هناك العديد من التقنيات الحديثة التي تم تطويرها لجعل تسلسل الحمض النووي أسهل وأرخص. من بينها، ينتج عن تنظيم الجينات أطول تسلسلات للقراءة وتتزايد أطوال القراءات يومًا بعد يوم. تم تطوير تقنية التعلم الآلي مثل Support Vector Machine (SVM) لمحاذاة تسلسلات الجينات هذه. واجهت كل تقنية بعض التحديات، لكن مواجهة التحديات المرغوبة تطرح بعض التحديات الجديدة على الجانب الآخر. لذلك، لا توجد أداة واحدة مثالية لكل عمل. تقنية SVM هي أداة تقويم جديدة تقوم بالمقايضة وتعمل بشكل أفضل من الجوانب المختلفة. بالنسبة للنموذج الذي يحتوي على أفضل خسارة في المولد، تم تحقيق متوسط دقة تحقق قصوى تبلغ 91.29٪. تنظيم الجينات باستخدام SVM يشبه الخريطة المصغرة التي تستغرق عدة مرات ذاكرة أكثر لفهرسة الجينوم الكامل للتسلسل المرجعي. تعد بيانات الخلية المفردة الهدف الرئيسي لـ SVM. من الواضح أنه سيساعد SVM والتقنيات المماثلة على التوافق بشكل أفضل مع عمليات الإدراج والحذف الطويلة لتنظيم الجين وحيد الخلية. يتم تشغيل بيانات الخلية المفردة مقابل تسلسل مرجعي معروف ومرجع اصطناعي تم إنشاؤه عشوائيًا.

INTRODUCTION

Gene regulation from single cell detection is very important for many cancers remedy. Even by detecting the mutations, tailored drug can be given to the affected patients tailored to the genetic makeup of their tumor or cancer cell [1]. Behind all these detections lies a common technique, which require Gene Alignment [2]. Alignment is a process where read sequences are aligned against a reference genome sequence of any particular species. However, any alignment first needed the reads to be mapped to the reference sequence. Means for each read there must be a position in the reference from where it was taken from [3]. For long reads which is prone to higher error rates like 15% to 20% and an average length of 10K base pare this read mapping becomes challenging actually [4]. Our goal was to develop a gene

regulation-mapping tool for long read which consume a feasible amount of time and gives a good mapping of the gene regulation from single cell data.

Motivation

The terms gene and regulation often used alternatively [6] to indicate a tool which align several portions of a read with several portions of a reference genome for certain reasons [7]. Nevertheless, in recent days, these words carry a slightly different meaning.

Gene seeks for the best fit with matched clusters in the reference genome based on some parameter; it just identifies most similar clusters that would help the aligner to align faster and easier with gene lambda. Figure 1 illustrates the concept.

On the other hand, aligner takes those mapping and align the read with the best suitable portion considering some insertion, deletion, mismatch if needed as mentioned in [8]. A small demonstration where aligner picks one of the possible alignments and tells the read best fitted here, and then stores this result in a specific format [9]. As generally, regulation is used for storing string type of data or information and then retrieve. Regulation is also known as digital tree, radix tree or prefix tree [10]. By formal definition, a regulation is a tree containing a collection of strings with a single node per common prefix.

Before defining the problem specifically, the first thing is to have a good literature review to ensure that we are not going to waste our time by rediscovering anything. In this case, there are more than 95 gene tools developed until now according to one of the benchmarking [11]. However, most of them are regulation. As it is stated before, regulation and gene are separated in recent years. Before that, these two words are used alternatively. On the other hand, all of them are not general-purpose tools; rather their dedication goes to some categories like DNA, RNA, bi-sulfite, miRNA etc. Here, some tools are presented which are best matches with our task. Later in this research, some basics are covered to understand the work properly as depicted by the figure 1 from source [12].

Problem Statement

For the greater understanding of genetic information measuring gene expression level in any particular species or in any particular individual is very important. For finding gene expression level one need to find the number of reads at different genes or target loci of the reference genome [13]. These counts of reads to each gene are then used to estimate expression levels. There are many usages of gene expression in human genetics. Some of are listed below.

- Classification of human tumors according to the gene level.
- Proper analysis and profiling of breast cancer. Ontological analysis for proper biological interpretation for gnomic data and results.
- Proper sub classification of cancer like Myeloid Leukemia.

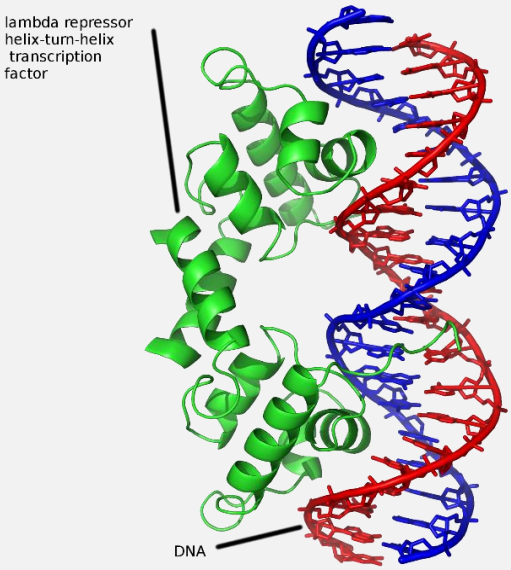

Figure 1. Picture of a gene lambda repressor helix-turn-helix transcription factor binding to its target location in DNA [12].

From the above list we can guess how important is to measure the gene expression level in any particular species or individual. In addition, most of the commonly used gene expression measuring tools need to align the read sequences short/long to a particular reference genome sequence. Moreover, we know that alignment is dependent on how well the reads are mapped to the reference genome sequence as given in [14]. This work is created utilizing the rationale-augmented convolutional neural network (CNN) [15]. In this research vehicle detection system from infrared images using YOLO (You Look Only Once) computational mechanism [16]. In this research, data clustering is an important machine-learning topic. It is useful for variety of applications one of them is image segmentation. [17]. Some of the many classification models are SVM (support vector machine), KNN (K- Nearest Neighbors), Decision tree, Logistic Regression and ANN (Artificial

ISSN: 1814-635X (print), ISSN:2521-3520 (online) Volume 33, Issue 3, 2022 DOI: http://doi.org/10.23851/mjs.v33i3.1069

Neural Network) back propagation. For this paper, we would consider different procedure and method of early detection of the glaucoma disease using the MATLAB Deep Convolutional Neural Network (DCNN) [18]. Changes in RNA structure could contribute to regulation of gene expression in a variety of ways. As RNA structure is vital for function, changes in RNA structure could either promote or inhibit gene expression as we show in Figure 2 from source [19].

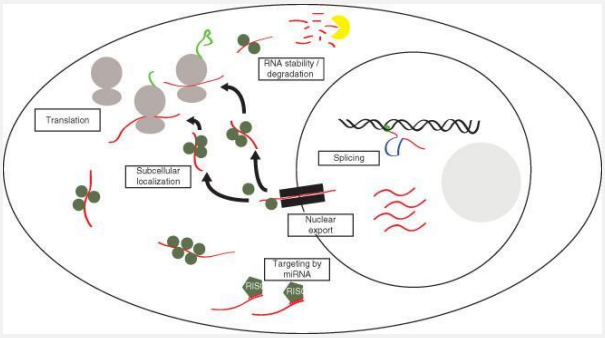

Figure 2. RNA structure is vital for function [19].

Algorithm and Solution

The research is based on machine learning based SVM gene regulation classifier with improved average maximum validation accuracy when it was trained on the augmented training sets. The results for the gene regulation are a significant improvement to the original baseline classifier without data augmentation and require further investigation. The inter-class, intra-class and centroid distances for the augmented dataset of gene regulation. While the intra-class distances follow, a similar trend to the ones are observe for the one-SVM per-class data augmentation strategy, the inter-class distances did severely decrease and are efficient for gene regulation detection and recognition.

Aim of Contribution

Gene regulation longs reads rather than short ones to a large genome sequence obviously presents an algorithmic challenge. Applications where such mapping and resequencing is being done a primary concern always remains how accurately it is done. Therefore, our goal is developing a gene regulation with SVM based mapping tool that can map features to the reference accurately and consuming a feasible amount of time for RNA, DNA and single cell genes. The research also aims to solve the gene RNA-binding protein, as they are a specific type of proteins mainly composed of DNA-binding domains and that is why having a specific or general inclination for either single or double stranded DNA. That means in many cases those proteins are supposed to bind to a specific site or location of the gene regulation DNA using machine learning techniques.

MATERIALS AND METHODOLOGIES

Basically, gene regulation expression is the physical method in which information from a gene is used in the synthesis of an operational gene product. These outcome products are mostly proteins, but in the non-protein coding genes such as transfer RNA (tRNA) or smaller nuclear RNA (snRNA) genes, the operational product is a functional RNA. Next Generation Sequencing (NGS) brought a tremendous revolution in the field of Bioinformatics. Gene DNA sequencing was very expensive before this revolution. Extracting information from gene sequences become very cheap and easy with the next generation sequencing technologies. Raw data kept as RNA or DNA format. The raw data file consists of short sequences, varying length from 30–60 base pair to 3K–20K bp1 each, taken from the random positions of the whole genome sequence. However, the genome sequence of all members of a single species are not same, rather every individual carries different genome than others. The difference between two genomes may be very few or may be very large. Based on this difference, some minor to major differences may be noticed in several things like behavior, skin, structure, color etc.. Individuals are born with mutation in their 20K gene. Figure 2 visualizes mutation clearly. However, identifying the mutations is a big issue for NGS data as the reads are short and repetitive as well as error occurs in reads. To retrieve the original genome sequence from these reads is a challenge. The latter achieved by doing the alignments and keeping in a special format called SVM. The more data, the more noise. The length of the reads would increase day by day. As petabytes of sequencing data is adding in the databases daily, it would be in no use if the data could not be processed efficiently.

Alignment is done for pointing structural, functional, evolutionary similar portions. An important part of alignment is mapping. An

alignment software does map first, but the concepts are divided into two parts in the recent past in a sense that, if mapping could be efficient, then the next steps of the process would be efficient. That is why efficient mapping technique in terms of memory, time and placement is become a challenging task. The dataset link: https://www.iccr-cancer.org/datasets/published-datasets/soft-tissue-bone

Figure 3 explains the systematic and pictorial view of gene regulation model. However, Table 1 highlights the SVM experiment modeling with number of parameters.

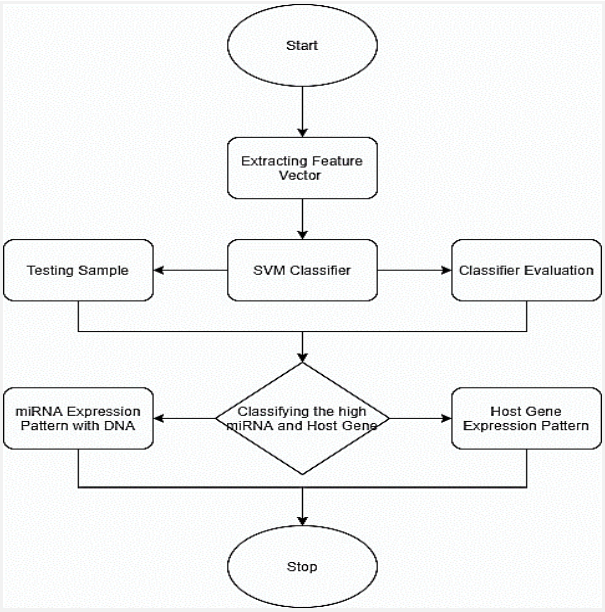

Figure 3. A systematic and pictorial view of gene regulation model.

Table 1. The SVM experiment modeling with number of parameters.

Data Partitioning	Percentage	K-Folds
Training	70%	10
Testing	20%	12
Validation	10%	10

We studied the effect of gene on the accuracy of prediction. To study this, we pruned the gene trained with 250 features being select for SVM classifier for classification on number of training samples. Then we calculated the test set accuracy of the pruned SVM models. It turned out that classification showed a great difference in accuracy compared to the full gene. We hypothesize the reason for this is that very high up in the gene the splitting nodes already confidently divide the training points into their class. However, the SVM algorithm continues to exhaust the features until the entire tree is constructed. By stopping the prediction early on in the gene, we do not suffer in accuracy, and improve the speed of prediction. Finally, we studied how much is the model overfits to the training samples, by calculating the training accuracy and comparing that with the test accuracy and in Figure 4 explain SVM based classifier trained on 100 epochs.

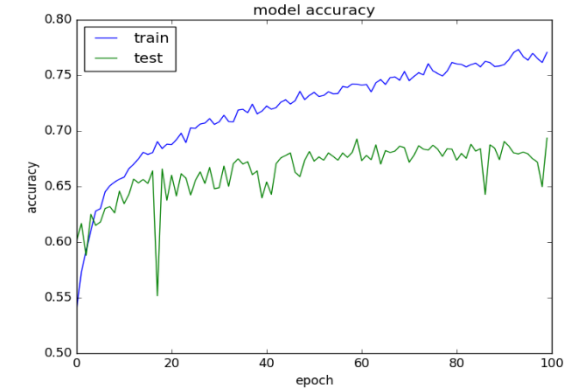

Figure 4. An SVM based classifier trained on 100 epochs.

RESULTS AND DISCUSSION

Research found that using the test accuracy is a good evaluation of the actual gene regulation system. Sometimes the system performs very well in the test data set, but good in practice. The reason behind this is that the environment we evaluate changes all the time: gene position, features, gene regulation, etc.. Accordingly, we choose to pick some significant boundaries dependent on our involvement with the genuine climate rather than simple test exactness. For instance, albeit the trials demonstrated the quantity of trees had no effect, we found this not to be the situation in real tests with differing foundations. Hence, we utilized a model prepared with different trees instead of one. The minimum generator loss model therefore outperformed the model with best inception accuracy, which asserts the impression that it is more suitable for the selection of the best gene regulation model. There was no particular formula being use for predicting the accuracy however the SVM self-evaluate the classification accuracy based on testing. Although our system is very time-sensitive, we found that it is not necessary to do optimization on one component while other components are not ready. We found it is best to develop quickly and do not profiling to discover the bottleneck and optimize it. This saves us a lot of time in making unnecessary optimization. Figure 5 explains the genome sequence of all members of a

single species in terms of DNA, RNA and Gene Regulation their percentage and number of final locations location, rather every individual carries different genome than others. Figure 6 represents the graphical plotting of gene regulation feature extraction with RNA and Gene Regulation in terms of improved average maximum validation accuracy when it was trained on the augmented training sets. It is perceived that gene regulation is working very fast with SVM in comparison with the RNA execution. Figure 7 explains the indexing memory consumption comparison between gene-index approach and RNA using SVM as it turned out that pruning the tree showed significant difference in accuracy compared to the full gene regulation as it takes less reference length as compared to RNA. Figure 8 depicts the right range mapping comparison among RNA, DNA and GENE for consuming a feasible amount of serial number of features for RNA, DNA and single cell genes for best possible results on regulation. Research fixed 250 features and 3 genes, and varied the number of training samples on 100 epochs. Research show evidence of extreme overfitting. The training error is consistently very low (mean 1.06%) and is much lower than the testing accuracy.

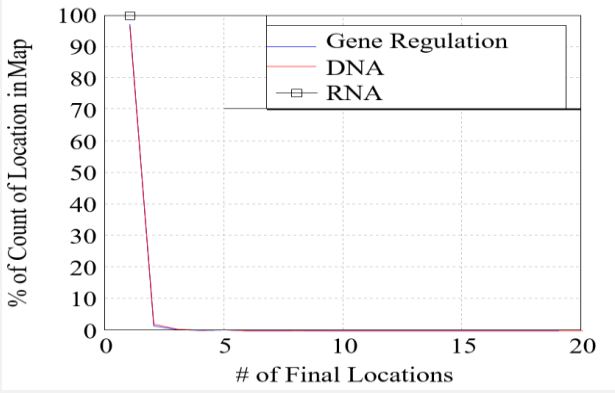

Figure 5. Count of Final Locations and Their Percentage among the whole Final Locations of Gene Regulation.

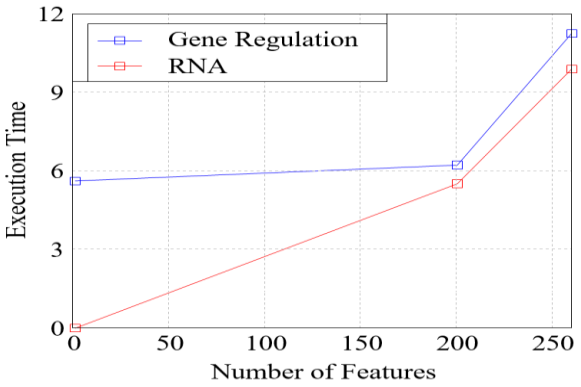

Figure 6. For error free data, enhanced gene-indexing working very fast with SVM. For noisy data, the time gap is reducing.

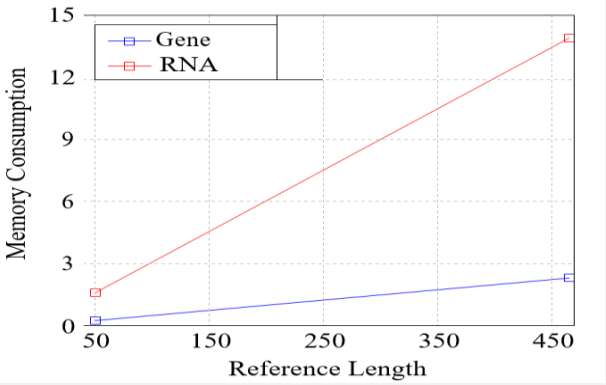

Figure 7. Reference Indexing Memory Consumption Comparison between Gene-index Approach and RNA using SVM.

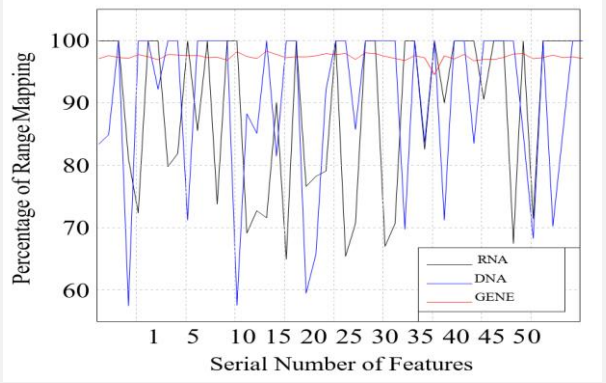

Figure 8. Percentage of Right Range Mapping Comparison Among RNA, DNA and Gene regulation.

A possible explanation for the observations on the maximum length of the synthetic sequences could come from the class sequence length distribution. While 95% of all original sequences are shorter than 200 frames, there are outlier sequences that are much longer and skew the arithmetic mean class sequence length of RNA, DNA and Gene Regulation. As the length of the synthetic samples is based on a gene RNA distribution that makes use

of these statistics, the generator might tend to produce longer sequences that help the classifier to better generalize to these rarer inputs. When clipped, this effect cannot be fully exercised by the generated samples. The smaller the number of training samples, the smaller is the number of original outliers, which increases the positive augmentation impact. Interestingly, the maximum validation accuracy is better for the multi-class gene regulation with single cell using SVM when the synthetic samples are clipped to 100 instead of 200 frames. In this case, the data augmentation emphasizes the inputs that tend to be shorter than the average. Nevertheless, these effects might be rather random statistic correlations and it is unlikely that the positive impact comes from the varying length of the samples alone. The base multi-class SVM had a positive effect on almost all groups of selected classes as well as on the complete dataset when used for data augmentation as comparison of accuracy is performed in Table 2.

Table 2. The information about the dataset being acquired.

Morphological Codes	Training Samples	Testing Sample	Validation samples
Benign	18000	8000	2500
Malignant	18000	8000	2500

This is a strong proof that the proposed SVM augmentation method is suitable for improving the classification on hand gesture action sequences. Modifications in the size of the critic architecture led to worse results. While the SVM-augmentation with the gene regulation critic scored a higher average best validation accuracy on the complete dataset, it performed worse when tested on the five-class set, showed some instability during training and generated sequences of lower visual quality. Thus, without the added gradient penalty term, the static SVM appears to be the superior and more stable architectural solution in Table 3 shows the comparison of the proposed method with state-of-the-arts techniques.

Table 3. The comparison of proposed method with existing technique.

Article	Technique	Accuracy
[20]	Random Forest	90.78%
[21]	Gaussian Naive Bayes	84.69%
[22]	Decision Tree	60 %
[22]	LDA	80%
[22]	CACGE	90%
Proposed	Support Vector Machine	91.29%

CONCLUSIONS

In this paper, the aim of study was to gene regulation with single cell interaction using advance machine learning based SVM classifier with data augmentation strategies were evaluated on a subset consisting of RNA, DNA and gene classes. The smaller dataset reduced training time and thus enabled a greater number of experiments in the limited amount of time. It was established that training all gene classes on a single cell is more effective for augmenting data than generating the samples of each class by a separate network. As the single class gene cell works with a greater number of training samples and different classes, it is more likely to capture the underlying structure of a gene regulation with RNA and DNA position but also the distinct differences between the action classes. Furthermore, the SVM model with the minimum generator loss had a bigger positive impact on classification accuracy than the model with the best inception accuracy, even though its synthetic sequences looked less realistic. Overall, the main objectives of the paper were successfully met. The established architecture was implemented and was shown to be effective for data augmentation. For the model with the best generator loss, an average maximum validation accuracy of 91.29% was achieved. In addition to this, the gene regulation was analyzed and compared to a range of alternatives based on suitable evaluation metrics.

REFERENCES

[1] Duncan, J.; Insana, M.; Ayache, N. Biomedical Imaging and Analysis In the Age of Sparsity, Big Data, and Deep Learning. Proc. IEEE 2020, 108. https://doi.org/10.1109/JPROC.2019.2956422

[2] Bojarski, M.; Del Testa, D.; Dworakowski, D.; Firner, B.; Flepp, B.; Goyal, P.; Jackel, L.D.; Monfort, M.; Muller, U.; Zhang, J.; et al. End to end learning for self-driving cars. arXiv 2016, arXiv:1604.07316.

[3] Huynh, B.Q.; Li, H.; Giger, M.L. Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. J. Med. Imaging 2016, 3, 034501. https://doi.org/10.1117/1.JMI.3.3.034501

[4] Spanhol, F.A.; Oliveira, L.S.; Petitjean, C.; Heutte, L. Breast cancer histopathological image classification using Convolutional Neural Networks. In Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, Canada, 24-29 July 2016; pp. 2560-2567. https://doi.org/10.1109/IJCNN.2016.7727519

[5] Han, Z.; Wei, B.; Zheng, Y.; Yin, Y.; Li, K.; Li, S. Breast cancer multi-classification from histopathological images with structured deep learning model. Sci. Rep. 2017, 7, 4172. https://doi.org/10.1038/s41598-017-04075-z

[6] Lévy, D.; Jain, A. Breast mass classification from mammograms using deep convolutional neural networks. arXiv 2016, arXiv:1612.00542.

[7] Liao, Q.; Ding, Y.; Jiang, Z.L.; Wang, X.; Zhang, C.; Zhang, Q. Multi-task deep convolutional neural network for cancer diagnosis. Neurocomputing 2019, 348, 66-73. https://doi.org/10.1016/j.neucom.2018.06.084

[8] Chapman, A. Digital Games as History: How Videogames Represent the Past and Offer Access to Historical Practice; Routledge Advances in Game Studies, Taylor & Francis: Abingdon, UK, 2016; pp. 185-185.

[9] Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genom. 2020, 21, 6. https://doi.org/10.1186/s12864-019-6413-7

[10] Zhang, Y.; Gong, D.W.; Cheng, J. Multi-objective particle swarm optimization approach for cost-based feature selection in classification. IEEE/ACM Trans. Comput. Biol. Bioinform. 2015, 14, 64-75. https://doi.org/10.1109/TCBB.2015.2476796

[11] Annavarapu, C.S.R.; Dara, S.; Banka, H. Cancer microarray data feature selection using multi-objective binary particle swarm optimization algorithm. EXCLI J. 2016, 15, 460.

[12] Alanni, R.; Hou, J.; Azzawi, H.; Xiang, Y. Deep gene selection method to select genes from microarray datasets for cancer classification. BMC Bioinform. 2019, 20, 608. https://doi.org/10.1186/s12859-019-3161-2

[13] Zhao, Z.; Morstatter, F.; Sharma, S.; Alelyani, S.; Anand, A.; Liu, H. Advancing feature selection research. ASU Feature Sel. Repos. 2010, 1-28. https://doi.org/10.1.1.642.5862

[14] Bolón-Canedo, V.; Sánchez-Marono, N.; Alonso-Betanzos, A.; Benítez, J.M.; Herrera, F. A review of microarray datasets and applied feature selection methods. Inf. Sci. 2014, 282, 111-135. https://doi.org/10.1016/j.ins.2014.05.042

[15] AHMED, Saadaldeen Rashid Ahmed; SONUÇ, Emrullah. Deepfake detection using rationale-augmented convolutional neural network. Applied Nanoscience, 2021, 1-9. https://doi.org/10.1007/s13204-021-02072-3

[16] MAHMOOD, Mohammed Thakir; AHMED, Saadaldeen Rashid Ahmed; AHMED, Mohammed Rashid Ahmed. Detection of vehicle with Infrared images in Road Traffic using YOLO computational mechanism. In: IOP Conference Series: Materials Science and Engineering. IOP Publishing, 2020. p. 022027. https://doi.org/10.1088/1757-899X/928/2/022027

[17] ABDULATEEF, Salwa Khalid; AHMED, Saadaldeen Rashid Ahmed; SALMAN, Mohanad Dawood. A Novel Food Image Segmentation Based on Homogeneity Test of K-Means Clustering. In: IOP Conference Series: Materials Science and Engineering. IOP Publishing, 2020. p. 032059. https://doi.org/10.1088/1757-899X/928/3/032059

[18] AHMED, Moahmmed Rashid, et al. An Expert System to Predict Eye Disorder Using Deep Convolutional Neural Network. Academic Platform Journal of Engineering and Science, 9.1: 47-52. https://doi.org/10.21541/apjes.741194

[19] Solem, A. C., Halvorsen, M., Ramos, S. B., & Laederach, A. (2015). The potential of the riboSNitch in personalized medicine. Wiley Interdisciplinary Reviews: RNA, 6(5), 517-532. https://doi.org/10.1002/wrna.1291

[20] Ni, Ying & Aghamirzaie, Delasa & Elmarakeby, Haitham & Collakova, Eva & Li, Song & Grene, Ruth & Heath, Lenwood. (2016). A Machine Learning Approach to Predict Gene Regulatory Networks in Seed Development in Arabidopsis. Frontiers in Plant Science. 7. https://doi.org/10.3389/fpls.2016.01936

[21] Kamel, Hajer & Al-Tuwaijari, Jamal. (2019). Cancer Classification Using Gaussian Naive Bayes Algorithm. 165-170. https://doi.org/10.1109/IEC47844.2019.8950650

[22] Alagukumar, S., and R. Lawrance. "Classification of microarray gene expression data using associative classification." 2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16). IEEE, 2016. https://doi.org/10.1109/ICCTIDE.2016.7725362

How to Cite

A. S. . Shaker and S. R. Ahmed, “Information Retrieval for Cancer Cell Detection Based on Advanced Machine Learning Techniques”, *Al-Mustansiriyah Journal of Science*, vol. 33, no. 3, pp. 20–26, Sep. 2022.