Journal of Advances in Science and Technology
Vol. 17, Issue No. 1, March-2020, ISSN 2230-9659

Review on Artificial Intelligence Based Techniques for Gene-Expression by Tumor RNA-SEQ Data

Pravin Vishnu Shinde[1]* Dr. Rajesh Deshmukh[2]

[1] PhD Student, Kalinga University, Raipur

[2] PhD Guide, Kalinga University, Raipur

Abstract – Human intelligence is needed or required byArtificial Intelligence (AI), or the development of computational algorithms that is successful in performing tasks that would've used human intelligence instead, helps in creating opportunities for improving our ideology and delivery of precision medication. Here, large-scale RNA-sequencing datasets in cancer are provided in this article as an overview of artificial intelligence approaches for of this analysis. Major Solutions for disentangle inter- and intra-tumor heterogeneity of transcript me profiles are displayed here for improving patient's management effectively. In this articlewe also outline the contributions that have been made by learning algorithms to the needs of cancerGenomics, from identifying rare cancer subtypes to personalizing therapeutic treatments.

Keywords – RNA Sequencing; Cancer Heterogeneity; Artificial Intelligence

----------X----------

1. INTRODUCTION

Cancer is normally referred to a collection of diseases which relates to growing of cells abnormally with invasive and metastatic features [1]. Cancer was accountable for more than 90 lakhs fatalities globally in 2018, alone. Roughly 20% of males 17% of females is expected to contract this dreaded diseasein their life at any point, and 13% of males and 10% of females and will eventually succumb to it [2]. Annually greater than 80 lakhs human souls perishes from cancer, as per statistics from WHO, resulting in approximately 13% of global deaths, focusing that cancer became one of most frightening diseases in today's world [1]. Most common cancers in 2018, recorded are colorectal cancer (860,000) and lung cancer (1.76 million deaths) .breast cancer (620,000), liver cancer (780,000), and Stomach cancer (780,000), ranked fourth, third and second out of most common types of cancers [2].

Chances of recovery increases if cancer diagnosed in early stages as damage done to critical organs is in recoverable state [3]. RNA sequence analysis in one such staging technique . Current advances in accuracy and efficiency of optimization algorithms and techniques of artificial intelligence have helped in analysis of human genomics. Depending on convolutional neural network (CNN) and binary particle swarm optimization with decision tree (BPSO-DT) [4] this research work put forth recent and new optimized deep learning schemes to group various kinds of cancer based on tumor RNA sequence (RNA-Seq) gene expression information. Types of cancer under this research investigates are breast invasive carcinoma (BRCA), kidney renal clear cell carcinoma (KIRC), uterine corpus endometrial carcinoma (UCEC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC).Proposed schemes constitutes of mainly three stages. Preprocessing is the first stage, for selecting only optimal characteristic utilizing BPSO-DT which at first optimize high-dimensional RNA-seq and consequently transform optimized RNA-Seq to 2D images. Augmentation forms second stage, that improves original dataset of 2086 samples to be more than a multiple of 5. Depending on achieving least impact on manipulating characteristics of images choosing of augmentations techniques are made. This stage assists in overcoming over fitting issues and prepares model to acquire good accuracy and precision. Deep CNN architecture forms third stage. This involves , an architecture of majorly two layers of convolution for characteristics extraction and 2 entirely joined layers is presented for categorizing five various types of cancer as per availability of data set's images. Cancer is normally referred to a collection of diseases which relates to growing of cells abnormally with invasive and metastatic features [1].

2. REVIEW OF RESEARCH WORKS

Related to field of tumor gene expression data,"this part carries out a survey on recent studies for deploying machine learning and deep learning approaches. For achieving important outcomes in a wider scope of medical image analyses and understanding tasks, researchers globally initiated in applying deep and machine learning tools. [3] indicate that Hsu et al. utilizes RNA sequencing information collected from The Cancer Genome Atlas (TCGA), and emphasizes on categorizing 33 different types of cancer effected patients. Five machine learning algorithms were presented by authors, namely, an artificial neural network (ANN), polynomial support vector machine (poly SVM), linear support vector machine (linear SVM), DT and KNN. Linear SVM is the best classifier in this study is depicted in result that, with a success ratio of about 95.8".

A new method was designed by Lyu and Haque [4] to "find out potential biomarkers for every malign kind of tumor. A new way has been providedwith allot of information on 33 prevalent cancer tumor types depending on pan- cancer atlas. For classifying tumor types they deployed a convolutional neural network and to discover top tumor genes from input, uses a visualization neural network techniques. So as to classify 33 cancer tumor types, high-dimensional RNA-Seq data was converted into 2-D images and is utilized as a convolutional neural network. Depending on concept of Guided Grad Cam, applied in each classes, they produced an important heat-map for entire genes structure. By applying a train/test split, suggested prototype system attained 95.59% of precision".

For development of a pan-cancer atlas to recognize 9,096 TCGA tumor samples representing 31 tumor types was undertaken by authors in [5]. They arbitrarily assigned 25% (approximately 2300 samples) into testing sets, 75% (approx. 6800 samples) of samples into training set and accordingly allocated samples from every types of tumor. For categorisation of non-sex-specific tumor classification, they excluded all tumor types that were sex-specific, such as, CESC, BRCA, PRAD, OV, TGCT, UCS and UCEC. For rest of kind of tumors, samples were segregated in two classes depending on patient's sex. Additionally Three types of tumor (KICH, CHOL and DLBC) were excluded because of small size of gender-specific samples. Authors deployed k-nearest neighbours (KNN) and genetic algorithm methods for iteratively generating subset of genes (characteristics) and subsequently using KNN techniques for testing accuracy and precision. Such techniques attains precision of 90% over 31 kinds of tumour and produces a set of top genes for all tumour kinds [6]–[9].

Deep learning schemes were also used to categorise genes of top tumors and identifies individual kinds of cancer. Authors primarily applies a stacked denoisingautoencoder (SDAE) in paper [6], [10], for extracting a high-level characteristics from high-dimensional gene expression profiles. To finalize whether a sample is a cancerous tumour or not, authors provides input to these features in a single-layer ANN network. Accuracy attained by deploying this method approaches to about 94%. Analysis and outcomes shows that these highly interactive cancer genes can be useful for detection of breast cancer in petients.

A semi-supervised deep learning techniques known as stacked sparse auto-encoder (SSAE) was presented by Xiao et al. [7] for grouping and to predict cancerous tumour using RNA-seq information. Suggested approach of SSAE-based method uses sparsity penalty term and pre-training greedy layer approach for capturing and extracting significant data from high-dimensional information and consequently classifying these samples. Presented SSAE technique was tired and tested on 3 public RNA-seq information sets of three kinds of cancers vis, breast invasive carcinoma (BRCA), stomach adenocarcinoma (STAD) and lung adenocarcinoma (LUAD). They compared prediction behaviour with many commonly employed classification schemes. These SSAE-dependent semi-supervised learning techniques attains best classifications of 99.89%, 96.23% and 98.15%, for LUAD, BRCA, and STAD datasets, respectively".

A novel strategy was demonstrated by Xiao et al. [8], that employs deep learning for an ensemble schemes which incorporates many various machine learning methods. Deep learning-based multi-model ensemble techniques under consideration were used for three public RNA-seq datasets on behalf of three kinds of cancers, breast invasive carcinoma (BRCA), stomach adenocarcinoma (STAD) and lung adenocarcinoma (LUAD). It provided an enhanced predictions of 99.20%, 98.78% and 98.41%, and for LUAD, STAD and BRCA datasets, respectively.

Deep learning schemes were recently being employed in many fields with great success [10]. But, their uses for analysis of high-throughput sequencing data still persists a challenging issue for research community because of the fact that these category of models are very well known to exhibit very well in big datasets with ample amount of samples available, contrary to this, opposite scenarios are typically found in the field of biomedical . In such research work, primarily approximation on application of deep learning for study and analysis of RNA-Seq gene expression profiles data is delivered. By applying a regularized linear model (standard LASSO) three public cancer-related databases are analyzed as baseline model, and 2 deep learning models which varies on characteristic selection technique employed prior to use of a deep neural network model. Outcome depicts that straightforward application of deep nets deployment which are available in public scientific

tools and in circumstances as explained within this research study is not sufficient to overtake simpler models such as LASSO. Hence , more complex and smarter way that incorporate before biological knowledge into assessment process of deep learning models may be importantly required so as to achieve better yields in terms of predictive behavior."

Detailed explanation about the various surveyed papers had been done. Some of the detection and classification methods to recognize the disease has been determined. Generally, the segmentation of the picture is done through clustering process. In addition, the detection and classification of the disease are done by computer aided methods and magnetic resonance imaging. As, S et al., 2019[11] proposed research on the development of convolution neural network (CNN) for categorization method of brain tumors in the T1 weigh contrast improved MRI pictures. The planned model comprised the main essential stages. One of the pre-processing of the pictures utilizing various image processing models and after that classification of the pre-processing pictures using image processing methods. In addition, the classification was done using CNN method. Experimental analysis was done on database 3064 pictures that comprises three kinds of brain tumor (glioma, meningioma, pituitary). They acquired the maximum testing accuracy up to 95%, total precision was 93.23% and recall value was 93.01% using CNN method. It was estimated from the experiment analysis that there was an improved accuracy on the database.

Shakeel, P. M. et. al., 2019 [12] implemented MLBPNN that was suggested using the infra-red sensor image technology. After that, computing the multi face behavior of the neural differentiating the fact that was incredibly destroyed when the complete system was destroyed into simple sub-networks. The various characteristics were eliminated through the fractal size approach and after that the main essential characteristics were selected through the multiple fractal size recognition method to decrease the rate of the complexity. The picture sensor was combined through the WIIS sensor that was created to transfer the tumor hot information to medical expert to display the welfare situation and for helping the regulation of the ultrasound consideration stage, mainly if that occurred at aging patients residing in the remote area.

Kaldera, H. N. T. K. et. al., 2019 [13] studied the issue of the whole automatic brain tumor classification and segmentation in MRI comprising Glioma and Meningioma kinds of the brain tumors were measured. This research presented faster-CNN model for segmenting issue along with decreased amount of calculations along with maximum accuracy rate. In this research, they utilized 218 pictures as the training group and networks demonstrated the accuracy about 100% in Meningioma and 87.56 in Glioma classification and the average confidence rate of about 94.7% in distributing of Meningioma tumors. The distributed tumor area were identified by the ground-truth observation and manual method using Neurologists.

Gumaei, A. et. al., 2019 [14] developed a crossbreed unique properties extracted technique along with regular EL equipment for the development of an exact brain tumor classification method. This technique by eliminating the characteristics from the brain pictures through the use of the hybrid feature extraction technique, and after that calculating the co-variance matrix (CM) of structures to plan the novel essential group of features using PCA model. Lastly, the (RELM) regular extended learning machine was utilized for the classification method of the kind of the cancer image. The planned model was implemented and compared with the group of the experiments that was analyzed on novel public database of the brain pictures. Experimental analysis demonstrated that have better performance comparable to current methods in term of classified accuracy up to 94.3% using random hold-out method.

Saraswathi, V et al., 2019 [15] presented research on the multiple class classification of the brain tumor in MR neurological pictures through randomized forest classification with three methods. In the planned model, the GLCM, LBP (Shape based and local binary pattern) characteristics were calculated. After that, PCA (principal component analysis) were utilized for the size deduction of the calculated feature vector (FV). The complete features were eliminated into smaller patches comprising 3X3 in localized window. The experiment analysis was done on brain tumor database comprising 3064 T1 weigh contrast improved pictures and comparative study of RF, RF-PCA and RF-PCA along with randomized selection. Experimental results showed that RF-PCA along with randomized selection performed better compared to other techniques along with testing accuracy was 8.8%and validated accuracy of about 85.4%.

3. REVIEW OF THE RESEARCH GAPS IDENTIFIED

With respect to pre-processing phase, this research work presented is novel in nature that constitute of design in deep learning architecture and optimization process. One of such concerned works provided in [4] made an identical contribution to that of ours. Main difference is shown in Table between work showcased in [4] and our put forth reesearch. Skeptic approach may take place here in [4] in this part of section and why not relate with other research work in [3], [5], [7], [8]. Major reason is that research presented in [4] employs identical techniques analogous to ours, that constitutes of transforming RNA sequence to images and subsequently deploying deep learning models, where as other related research applies various

schemes that will be partial to compare our study with those. Moreover, choosen dataset [9] which is applied in this work in recently presented and published in May 2018, therefore it remains an open chance to research the proposed model on such newly published information. Suggested architecture in these tasks had decreased training information however applying adopted augmentation methods provided a better overall testing precision and accuracy with 96.16% overtaking related research with respect to overall testing accuracy for 5 categories. Moreover, this research does not uses any kind of thresholding values and consider training of entire dataset, where as related work had dropped few of thresholding methods on information. Furthermore, suggested deep learning architecture is power in complexity, as it constitutes of 14 layers only. Whereas related research consist of 23 layers with a large amount of neurons in completely joined layers with 1024 FC layer number 2 and 36864 neurons in FC layer number 1 that can reflect on time of training on hardware., Related work attains better testing precision and accuracies On other hand, for LUAD and BRCA, where as ours obtained better testing accuracies for UCEC, LUSC and KIRC [16]-[18].

4. CONCLUSION

Cancer was accountable for more than 90 lakhs fatalities globally in 2018, alone. Current advances in accuracy and efficiency of optimization algorithms and artificial intelligence techniques and have helped in analysis of human genomics. Augmentation forms second stage, that improves original dataset of 2086 samples to be more than a multiple of 5. Annually greater than 80 lakhs human souls perishes from cancer, as per statistics from WHO, resulting in approximately 13% of global deaths, focusing that cancer became one of most frightening diseases in today's world. Most common cancers in 2018, recorded are colorectal cancer (860,000) and lung cancer (1.76 million deaths). Proposed schemes constitutes of mainly three stages. This stage assists in overcoming over fitting issues and prepares model to acquire good accuracy and precession.

REFERENCES

[1]. Y. H. Zhang (2017). ''Identifying and analyzing different cancer subtypes using RNA-seq data of blood platelets,'' Oncotarget, vol. 8, no. 50, pp. 87494–87511.

[2]. F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal (2018). ''Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries,'' CA, A Cancer J. Clinicians, vol. 68, no. 6, pp. 394–424.

[3]. Y.-H. Hsu and D. Si (2018). ''Cancer type prediction and classification based on RNA-sequencing data,'' in Proc. 40thAnnu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC), pp. 5374–5377.

[4]. B. Lyu and A. Haque (2018). ''Deep learning based tumor type classification using gene expression data,'' in Proc. ACM Int. Conf. Bioinf., Comput. Biol., Health Informat. (BCB), pp. 89–96.

[5]. P. Danaee, R. Ghaeini, and D. A. Hendrix (2016). ''A deep learning approach for cancer detection and relevant gene identification,'' in Proc. Pacific Symp. Biocomputing., vol. 22, pp. 219–229

[6]. Y. Xiao, J. Wu, Z. Lin, and X. Zhao (2018). ''A semi-supervised deep learning method based on stacked sparse auto-encoder for cancer prediction using RNA-seq data,'' Comput. Methods Programs Biomed., vol. 166, pp. 99–105.

[7]. Y. Xiao, J. Wu, Z. Lin, and X. Zhao (2018). ''A deep learning-based multi-model ensemble method for cancer prediction,'' Comput. Methods Programs Biomed., vol. 153, pp. 1–9.

[8]. K. N. C. Ferles and Y. Papanikolaou (2018). ''Cancer types: RNA sequencing values from tumor samples/tissues,''. Distributed by Mendeley. [Online]. Available: https://data.mendeley.com/datasets/sf5n64hydt/1.

[9]. Urda, D., Montes-Torres, J., Moreno, F., Franco, L. and Jerez, J.M. (2017) Deep learning to analyze RNA-seq gene expression data. In International work-conference on artificial neural networks (pp. 50-59). Springer, Cham.

[10]. Das, S., Aranya, O. R. R. & Labiba, N. N. (2019, May). Brain Tumor Classification Using Convolutional Neural Network. In 2019 1st International Conference on Advances in Science, Engineering and Robo tics Technology (ICASERT) (pp. 1 5). IEEE.

[11]. Shakeel, P. M., Tobely, T. E. E., Al Feel, H., Manogaran, G. and Baskar, S. (2019), "Neural network based brain tumor detection using wireless infrared imaging sensor," IEEE Access, 7, 5577 5588.

[12]. Kaldera, H. N. T. K., Gunasekara, S. R. and Dissanayake, M. B. (2019), "Brain tumor Classification and Segmentation using Faster R CNN,"In 2019 Advances in

Science and Engineering Technology International Conferences (ASET) (pp. 1-6). IEEE.

[13]. Gumaei, A., Hassan, M. M., Hassan, M. R., Alelaiwi, A. and Fortino, G. (2019), "A hybrid feature extraction method with regularized extreme learning machine for brain tumor classification," IEEE Access, 7, pp. 36266-36273.

[14]. Saraswathi, V. and Gupta, D. (2019), " Classification of Brain Tumorusing PCA-RF in MR Neurological Images," In 2019 11th International Conference on Communication Systems & Networks (COMSNETS) (pp. 440-443). IEEE.

[15]. E. Gibson, W. Li, C. Sudre, L. Fidon, D. I. Shakir, G. Wang,Z. Eaton-Rosen, R. Gray, T. Doel, Y. Hu, T. Whyntie, P. Nachev, M. Modat, D. C. Barratt, S. Ourselin, M. J. Cardoso, and T. Vercauteren (2018). ''NiftyNet: A deep-learning platform for medical imaging,'' Comput. Methods Programs Biomed., vol. 158, pp. 113–122.

[16]. S. Liu, Y. Wang, X. Yang, B. Lei, L. Liu, S. X. Li, D. Ni, and T. Wang (2019). ''Deep learning in medical ultrasound analysis: A review,'' Engineering, vol. 5, no. 2, pp. 261–275.

[17]. Y. LeCun, Y. Bengio, and G. Hinton (2015). ''Deep learning,'' Nature, Vol. 521, No. 7553, pp. 436–444.

[18]. G. Eraslan, Ž. Avsec, J. Gagneur, and F. J. Theis (2019). ''Deep learning: New computational modelling techniques for genomics,'' Nature Rev. Genet., vol. 20, no. 7, pp. 389–403.

[19]. A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis (2018). ''Deep learning for computer vision: A brief review,'' Comput. Intell. Neurosci., Vol. 2018, Art. No. 7068349.

[20]. J. Riordon, D. Sovilj, S. Sanner, D. Sinton, and E. W. K. Young (2019). ''Deep learning with microfluidics for biotechnology,'' Trends Biotechnol., vol. 37, no. 3, pp. 310–324.

[21]. J. You, R. D. Mcleod, and P. Hu (2019). ''Predicting drug-target interaction network using deep learning model,'' Comput. Biol. Chem., vol. 80, pp. 90–101.

[22]. K. Jaganathan, S. Kyriazopoulou Panagiotopoulou, J. F. Mcrae, S. F. Darbandi, D. Knowles, Y. I. Li, J. A. Kosmicki, J. Arbelaez, W. Cui, G. B. Schwartz, E. D. Chow, E. Kanterakis, H. Gao, A. Kia, S. Batzoglou, S. J. Sanders, and K. K.-H. Farh (2019). ''Predicting splicing from primary sequence with deep learning,'' Cell, vol. 176, no. 3, pp. 535.e24–548.e24.

[23]. C. Cao, F. Liu, H. Tan, D. Song, W. Shu, W. Li, Y. Zhou, X. Bo, and Z. Xie (2018). ''Deep learning and its applications in biomedicine,'' Genomics, Proteomics Bioinf., vol. 16, no. 1, pp. 17–32.

[24]. D. Ciresan, U. Meier, and J. Schmidhuber (2012). ''Multi-column deep neural networks for image classification,'' in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 3642–3649.

Corresponding Author

Pravin Vishnu Shinde*

PhD Student, Kalinga University, Raipur