Novel Architecture for carcinoma prediction using convolution neural network

Abdulhamza A. Abdulkarim [(1)]
Al-Mamoon University College, Department of Computer Engineering Techniques

Wameedh Riyadh Abdul-Adheem [(2)]
Al-Mamoon University College, Department of electrical power Techniques

ABSTRACT

Neural networks are strong techniques commonly employed to create microarray data models of cancer prediction. We examine the most recent models suggested to emphasize neural networks' contributions in cancer prediction using data on gene expression. In 2013-2018, we identified documents published in science databases with keywords such as cancer classification, analytics of the cancer, cancer prediction, cancer clustering and micro-array information. Studies demonstrate that neural network methods have been used in cancer prediction, kind of cancer or risk of survival in the filtering of (data-engineering) genetic expressions or in collecting unlabeled samples. The study also covers some practical problems that may be taken into account while developing a brain cancer prediction model based on the network. The results show that its overall architecture influences the functionality of the neural network. However, decisions are taken with the use of trail-and-error approaches regarding the amount of hidden layers, neurons, hypermeters and algorithms.

1.1 Introduction

Cancer is the second greatest world cause of death, with an average of 1 in 6 cancer deaths. Substantial research has been carried out on cancer diagnosis and treatment techniques in attempt to decrease its impact on human health. The primary aim in the prediction of cancer is to classify tumor kinds and to discover indicators for each cancer, so that a teaching machine can be developed for determining at the earliest stage a metastatic tumor type and cancer. The cancer susceptibility, recurrence and prognosis are crucial to cancer predictions. The need for new machine learning methods for discovering new bio-markers is an important driver in many clinical and interpretation applications thanks to an increasing knowledge of precision and early detection techniques, including a variety of screens for detections with a sensitivity of around 70-80 per cent. One of the most often used techniques to analyse genetic disorders is micro-array technology. The standardized microarray dataset comprises a few 100 samples and thousands of gene expressions. Each expression quantifies the activitative level of the gene inside a particular tissue, which offers good insight into the disease pathogen, allowing improved diagnosis and predictions of future samples, by comparing the genes expressed within aberrant malignant cells with those in normal tissues.

Breast cancer is most likely to occur among all forms of cancer in women. After lung and brain cancer, breast cancer has the second-highest death rate, yet only around 30 percent of newly diagnosed cases have breast cancer. Advancing the fight against cancer requires early detection which can only be possible with an efficient detection system. Techniques have been developed to detect breast cancer, including medical image processing and digital pathology. Images are acquired by histopathology, which generally includes biopsy of the affected tissue. Tissues affected by the tumor are extracted by the pathologist and stained by H& E, which is the combination of histological stains called hematoxylin and eosin, after which it is examined under a microscope for cancerous cells by finding malignant features in cellular structures such as nuclei. These microscopic images can be collected and used for developing computer-aided detection systems. Manual detection is a tedious, tiring task and most likely to comprise human error, as most parts of the cell are frequently part of irregular random and arbitrary visual angles. The goal is to identify whether a tumor is benign or of a malignant in nature, as malignant tumors are cancerous and should be treated as soon as possible to reduce and prevent further complications. It can be solved by different machine techniques, in other words, a problem in the binary classification. It has been shown in the past that machine learning algorithms perform better than a human pathologist. A majority of scholars have found that medical image processing using machine learning provides more accurate results as compared to the objective diagnosis given by a pathologist. A study in Europe has been conducted by Phillips in which a set of algorithms along with breast images provided more accurate detection. This finding is also evidence that using high-resolution images and better algorithms will improve the performance and accuracy of cancer detection.

1.2 Neural Network methods

Neural networks are strong methods of machine learning, extensively used for learning data representation at several abstraction levels. This information is helpful in various applications like rebuilding, classification, Grouping and acknowledgement. The resultant categorization, grouping or statistical analysis on the data use predictive models such as cancer prediction models.

On the basis of our study of the latest cancer prediction model studies, the neuronal network techniques now available are classified as having functionality:

(i) filtering (preprocessing) methods
(ii) predicting (classification) methods
(iii) clustering methods.

Neural filtering network methods are utilized for extracting representations that best describe genetic expressions, without directly taking the objective prediction into consideration. Prediction and grouping alternatively addresses extract depictions which enhance predictive accuracy or divide the genes or samples into groups based on mutual similarity.

In this study, we look at the newest model neural network prediction by giving the instruments and designs utilized for preparing the data. We also offer a short discussion to emphasize some crucial questions that may be taken into account while constructing new models of cancer prediction. The presentation of neural network models specially built for the forecast of cancer using gene expression data differentiates this study from others.

1.3 Deep Learning Approach

Deep Learning is a multi-stage model of a neural network that is excellent in the field of large data learning. Like other machine training techniques, DL is a training stage in order to estimate network parameters from a given training dataset and to test the network to prevent the output of new input data. The creation of the DL model for a better prevision of cancer type preciseity and new interpretability was possible by the increase in the transcriptome profiles of tumor samples.

2. Literature Review

There are various ways to detect breast cancer including Mammography, Magnetic Resonance Imaging (MRI) Scans, Computed Tomography (CT) Scans, Ultrasound, and Nuclear Imaging. Although, none of these aforementioned techniques gives a completely correct prediction of cancer. Tissue-based diagnosis is mainly done with a staining methodology. In this procedure elements of tissues are colored by some staining element, usually hematoxylin and eosin (H&E). Cell structures, types, and other foreign elements are stained accordingly, and are easily visible under high resolution. Pathologists then examine the slide of stained tissues under a microscope or using high-resolution images taken from the camera. For detection of tumors, a histopathology test is essential. It is an old method used to predict invasive cancer cells from H&E stained tissues. There are various shortcomings for this procedure as it involves intra-observer variation, cancer cells and tissues can also have multiple appearances, and many other figures in cells have the same hyperchromatic features, which make identification difficult. The choice of area is also a factor as the process is done only on a small area of tissue, so the chosen area should be in the tumor periphery. Deep learning methods can be used to solve the difficulties described above. Deep learning is a prominent subset of machine learning technology, which is inspired to examine unstructured patterns in the human brain. Deep learning models have a high success rate because they train on hierarchical representations. Moreover, they can extract and organize different features and hence do not require any prior domain knowledge. On the other hand, trivial methods need rigorous feature engineering to obtain features, which involves domain expertise. Many deep learning methods have been proposed to predict the class of tumor.

There are various methods and manual networks proposed by various scholars to classify breast cancer other than the predesigned networks stated above. For example, Artificial Neural Networks depend upon MLE (Maximum Likelihood Estimation). The GRU SVM is the ML method in conjunction with a kind of neural recurring (RNN) and recurrent control unit (GRU) with a carrying vector machine (SVM). In deep learning algorithms, a series of tasks are implemented. The first step is image preprocessing which is required to convert data into the format in which it can directly be input to the network. This step involves multiple channeling of images, then segmentation is done (only if required, e.g. if there is a need to separate regions of interest from the background or omit parts that are not needed for training). On this stage, data is ready to be used in training, either in a supervised or an unsupervised manner. The next step is feature extraction. Features represent the visual content of the histopathology image. In the case of supervised feature extraction, features are known and different strategies are applied to find them, but in case of unsupervised feature extraction methods, features are not known and acquired implicitly in proposed solutions through the Convolutional Neural Network (CNN). The last step is classification, which places an image into the respective class (benign or malignant) and can be done using SVM (support vector machine) or with a fully connected layer using an activation function such as Soft max.

2.1 Neural Networks

Neural networks are powerful toolkits for resolving difficult non-linear problems and for revealing universal input/output mappings [46]. Consider a fully-connected Multi-Layer Perceptron network (MLP) with L layers: input layer, hidden layer sequence and output layer, for a better understanding of the notion. The layers are indexed according $l = \{0, \ldots, L-1\}$ and each layer has a number of neurons equal to nl. We will denote each input training example x as I x I I I ( ) [, , ...,] = 1 2 n0 and its output O x O O O ( ) [, , ...,] = 1 2 nL 1 . The network is trained to enter inputs in each neuron such that the activation value is determined. The activity levels of neurons on the output layer are calculated and collected to achieve O (x). To calculate the difference between O(x) and the required

output form, a default value is utilized. This is a standard feature. The objective will be changed using a back-propagation approach to modify weights to an optimum error value by propagating error derivatives across the network. One of the most often utilized designs and training Algorithms are the mentioned feed forward layered networks and feedback mechanism. They are extensively utilized with gene expression data. We focus on them. Prieto provided an introduction of neural network modelling, simulation and implementation, including examples of models utilized for solving issues in the actual world. In the next sections we will examine prior studies that provide models for neural network prediction of cancer. These models employ MLP, neural networks or generative opposite network designs to learn the characteristics of gene expression. Although the amount of neurons and topologies in the network is different, all techniques employ the same training-algorithm as the above.

2.2 Cancer prediction models

Cancer prediction models consist of one or more techniques that work together towards a high degree of prediction accuracy. Statistical methods and machine learning technologies were widely used to build cancer prediction models, enabling physicians to make exact forecasts, tailored treatments, and decrease patient expenses. The accuracy of cancer prediction models depends on the input data. Genes are big and incorporate noise that decreases the classification accuracy. Expressions of the genes. The information is also spatially organized and so can improve the ability to discriminate in the model. The performance of prediction model prediction may be evaluated using several steps, such as accurateness, recall, specificity, precision, prediction of negative rate, Matthew correlation factor and F1. A further measure displays the true positive rate against false positive rate (Sensitivity) The receptor operational characteristic curve (ROC) (specificity). It shows the likelihood to categorize a positive case randomly over a randomly picked negative one.

2.3 Convolutional neural network

CNN is a modified deep neural net type which depends upon nearby pixels being correlated. It employs randomly determined input patches at the beginning and changes them in the training phase. When training is performed, the network utilizes these changed patches to predict and confirm the testing outcome. The success of the picture categorization challenge was obtained using evolutionary neural networks as the defined nature of CNN matches the data point distribution in the image. As a result, many image processing tasks adapt CNN for automatic feature extraction. CNN is frequently used for image segmentation and medical image processing as well.

Two major forms of change are in the CNN architecture. The first is a convolution, in which pixels with a filter or kernel are convolved. The product between the picture patch and kernel is presented in this phase. Depending on the network the width and height of the filters may be defined and the filter's depth is equal to the input depth. The second essential transformation is subsampling, which may be utilized by various kinds and according to specific requirements (max pooling, min pooling, and average pooling). The user may select the size of the pooling filter and is typically used in strange numbers. The pooling layer is responsible for reducing the size of the data and is very helpful in reducing overlap. The output can be sent to a totally connected layer for effective classification by utilizing a combination of convolution and pooling layers. The whole process visualization is provided Fig 2.1

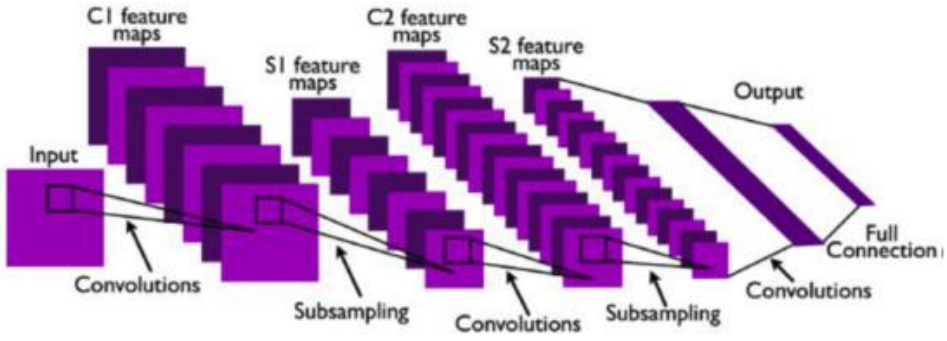


Fig 2.1 CNN Architecture

3. Research Methodology

3.1 Neural network-based cancer prediction models

A thorough search of the neural web-based cancer prediction was done using Googles scholar and two other electronic databases, PubMed and Scopus. A search was conducted using such words as 'Neural Networks' and 'Krebs Classification,' 'Neuro network,' 'Neuronal Network,' 'Gene Expression,' 'Neural Network' and 'Microarray,' and 'Neuronal Networking AND Cancer Predictions.' The approach includes only articles published between 2013-18 that were available publicly for free, including one or many neural network

models. The selected publications included categorization, discovery, survival prediction and statistical analysis approaches. For papers, the inputs of imaging or text (record) were removed.

3.2 Dataset and preprocessing

Most automated cancer prediction and clustering study investigations have employed available publically available datasets such as TCGA, NCBI Gene Expression Omnibus (GEO) and biomedical database Kentridge. Different methods were used by selecting a selection of genes to decrease the dimensionality of the gene expressions. Choosing a subset of related genes as described for other studies or data repositories like Online Mendel's Man heritage (OMIM) is another simple method of selecting features, the same preprocess methods as Chisquare selection features, selecting top 10 gene signatures related to lung cancer and combining them with clinical data in the T and N stages. The expression of the randomly determined number of genes is set to zero rather than choose a subset of genes. Xiao used DESeq to identify a set of genes, most informative genes, based on their reading figures.

3.3 Neural network architecture

Our analysis of recent research reveals that neural network methods are used in cancer prediction models for:

(i) Filtering or decreasing their dimensionality by eliminating genetic expressions. Statistical techniques and classification and clustering tools, such as decision books, K-Nearest Neighbor (KNN) and Self-Organization Maps (SOM) are the resultant features in fig 3.1.

(ii) Processes of prediction, which extract features that increase prediction accuracy (classification). This combines the lowering of dimensions of the same learning system with forecasting objectives.

(iii) Cluster techniques which, based on their resemblance, split gene expressions or samples.

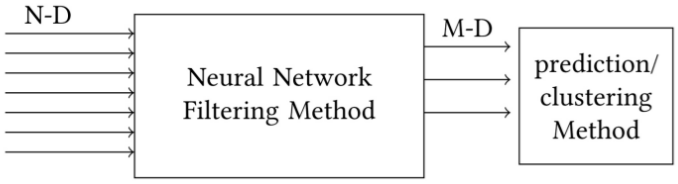


Fig 3.1 Neural networks for filtering the gene expressions in cancer prediction models.

3.4 Neural network clustering methods in cancer prediction

Based on their feature similarity Clustering is an unattended learning approach, which separates incoming data samples into groups. The classic model-based clustering approaches are neural networks, SOM which is often used with gene expression data, in particular. SOM is a one-layer neural network projecting massive data inputs onto a grid. The neurons on the output layer of a SOM are organized into a two or three-dimensional map, where each neuron represents a group and related clusters are located near each other via simple neighborhood functions. SOM connects every neuron in its output to the reference vector learned during the training process and transfers the neuron to every data point with the nearest reference vector. SOM learns with unlabeled input and with no background mechanism using a purely unmonitored method. Its preciseity may be assessed by several evaluation matrices, such as the Rand Index, one of the most often used gene expression data clustering methods.

$$RI = \frac{TP + TN}{TP + TN + FP + FN}$$

3.5 CNN with matrix input and 1D kernels

The third model is the 2D-Hybrid-CNN, influenced by the Resnet towers and simple 1D-CNN. 2-D inputs with the fundamental principles of 1-D convolution are expected to be used. Two 1D kernels are divided in this model across the inputs, one slides horizontally across the 2D input and the other with the column size. The 2 1D kernel outputs are then sent through the pool layer to the FC and forecast levels. Like resnet modules, we anticipate this design to capture unstructured features in the input genes' expression.

3.6 Implementation of 2D-3Layer-CNN

In order to provide a fair side-by-side comparison between the CNN model created in this research, DL platform was introduced and labelled 2D-3LayerCNN. This model comprises three converting modules which are coupled in cascade in each case by batch standardization, activation function (AF) and max pooling. The end module is supplied with two FC layers, and soft max is eventually employed for forecasting 33 different forms of cancer.

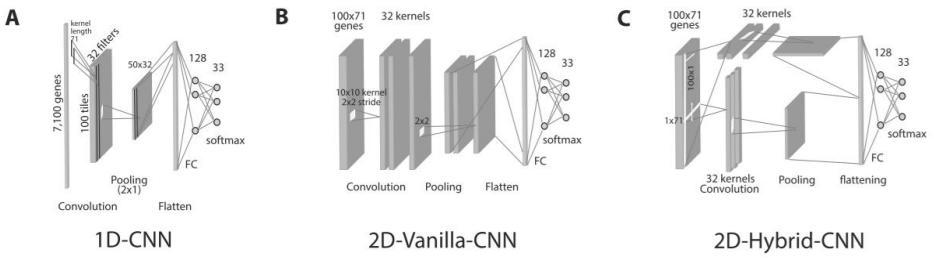


Fig. 3.2 Illustration of three CNN models. a 1D-CNN with input as a vector format with 7100 genes. b 2D-Vanilla-CNN, with an input reformatted as a 100 × 71 matrix, and one convolution layer. c 2D-Hybrid-CNN, similar input as in (b) but with two parallel convolution layers, vertical and horizontal, as in (a)

4. Result and Discussion

4.1 Model construction, hyper parameter selection and training

Keras DL platform has integrated all three models. 1D-CNN is input for a vector 1D following the alphabet order of the gene symbol, and inputs have been rearranged to 100 rows with 71 column matrix on both 2d-VanillaCNN and 3D Hybrid-CNN models. The search grid technology specifies the number and size of kernels, the kernels step as well as the number of the FC layer parameters. In addition, Categorical Cross, Entropy as loss function, Categorical accuracy as training measure and Adam Optimizer have been adopted for all three CNN models. In the case of a study of accuracy in category four successive time frames, the age and the size of the lot were respectively 50 and 128 and the early end of their learning is patient = 4 correspondingly. Finally, ReLU was used for the finishing layer of all models as AF and soft Max. Predictive level. Predictive layer.

Initially, all three CNN models with all 10340 tumor samples were trained. In order to test training, validation and strength against overfitting, Loss functions for three models with a division of 80–20 percent have been studied and ~0 loss after 10 years has been observed (who validation loss is about 0.10 and no overfitting is evident). The model has been equally trained and tested. This model is slower to converge than the three models discussed in this paper. A cross-validation has been done five times (due to the time restriction). To reduce the differences in the classifications precision, averaged and standard differences were provided for all models to minimize the bias caused by the stochastic dependent nature of the neural system in training.

4.2 Predicting cancer types without the influence of tissue of origin

We have included a new label into a prediction layer to take into consideration all normal samples, in order to take the influence of the origin tissue into the model (regardless of their original tissue type designation). With the aim of creating a robust cancer type prediction, the 34th node of the prediction layers removes the trace of origin tissue from cancer sample. All three models have reworked in accordance with the identical designs with 33 tumor-class nodes plus one normal sample node. Like models with just 33 kinds of cancer, we had a consistent learning curve with a division of 80–20 percent for training and validations, which we converged to almost 0 losses over 10 periods without any apparent overfitting.

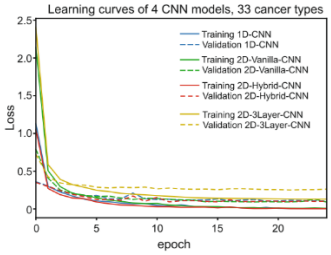


Fig 4.1 Learning curve for all three CNN models.

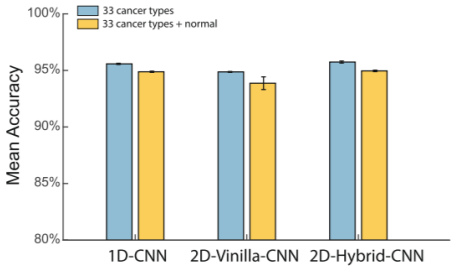
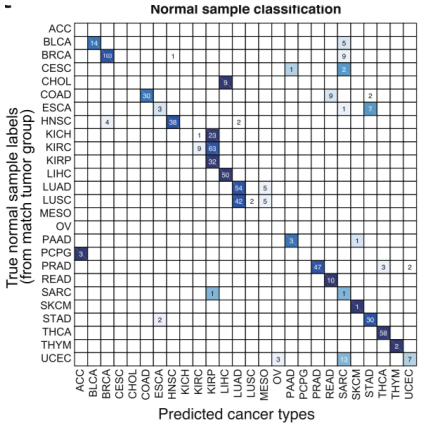
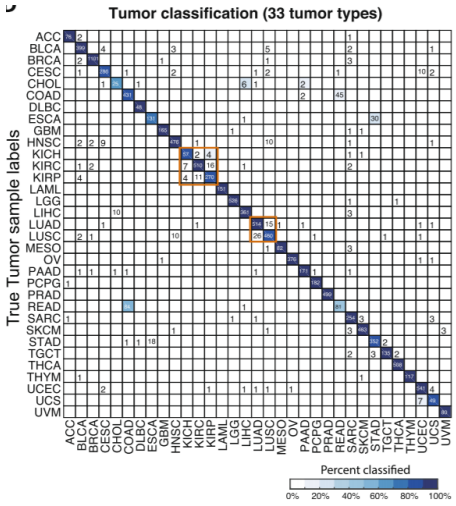


Fig 4.2 Micro-averaged accuracy of three CNN models



4.3 Confusion matrix of normal samples prediction from 1D-CNN model



4.4 Confusion matrix of 1D-CNN model, Tumor Classification

4. 3 Discussion

This article addressed many key aspects with the aim of improving the accuracy of our forecast and interpretation. In particular, Three CNN architectures provided research on the appropriate architecture for non-structured gene expressions for the prediction of cancer types. The result was improved by 1D-CNN and 2D-Hybrid-CNN compared to 95% (95.6 percent). 2D-Vanilla-CNN comprises a single layer and 32 kernels, however it is a more complicated architecture that is composed by many DL modules 2 D-3LayerCNN. We emphasize several basic design features, the number of parameters for each model and loss value after training, testing and running time behind every proposed models.

- 1D-CNN is far simpler in literature than the other models. It does not need the arrangement of inputs in a certain order and just comprises one convolutionary layer. This streamlined design leads to a considerable decrease in the number of hyper parameters to be

calculated during training (from 26 million to around 200,000). Due to the difficulties and expensive costs in gathering big genome data in DL applications in genomic studios, that is very desirable.

- 2D-Vanilla-CNN contains a million hyper parameters that are much higher than 1D-CNN parameters. When a kernel step was set to be 1:1, the model proved more difficult to converge. It also increased accuracy with the capacity to collect additional global properties via sliding two different convolution kernels over the two orthogonal dimensions.

5. Conclusion

This study examined the latest cancer prediction models based on the neural network and analytical techniques of gene expression. In this review, a few typical designs, data bases and the accuracy of each proposed model were given. Analysis of the articles studied showed that the methods of a neural network may be used as filters, predictors and clusters. The filtering of neural networks is used to decrease and eliminate noise from the dimension of gene expressions. MLP and neuronal network classification methods were used with binary and multi-classification problems while testing and error determined the number of hidden network layers and nodes. The most effective way of achieving high cluster accuracy is by a hybrid technique incorporating both clustering and projective clustering. Deciding on the architecture of the cancer prediction is one of the problems for designers because there is no set guideline for ensuring high accuracy in prediction. Most research has calculated the number of hidden layers and neurons based on trail and error.

However, the neural network's role influences its overall architecture. This study has shown. This paper summed up the latest techniques and related designs in the neural network. We also highlighted a number of crucial aspects that have to be taken into account while building a neural network overlay and class imbalance model. In future, neural network based techniques will be more powerful by picking various network parameters or merging two or more of the ways given.

References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. CA Cancer J Clin. 2018; 68(1):7–30.
2. Cohen JD, Li L, Wang Y, Thoburn C, Afsari B, Danilova L, Douville C, Javed AA, Wong F, Mattox A, et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. Science. 2018;359(6378): 926–30.
3. Haendel MA, Chute CG, Robinson PN. Classification, ontology, and precision medicine. N Engl J Med. 2018;379(15):1452–62.
4. Phallen J, Sausen M, Adleff V, Leal A, Hruban C, White J, Anagnostou V, Fiksel J, Cristiano S, Papp E, et al. Direct detection of early-stage cancers using circulating tumor DNA. Sci Transl Med. 2017;9(403):eaan2415.
5. Schiffman JD, Fisher PG, Gibbs P. Early detection of cancer: past, present, and future. In: Am Soc Clin Oncol Educ Book: American Society of Clinical Oncology; 2015. p. 57–65.
6. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–44.
7. Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, Staudt LM. Toward a shared vision for Cancer genomic data. N Engl J Med. 2016; 375(12):1109–12.
8. Ahn T, Goo T, Lee C-h, Kim S, Han K, Park S, Park T. Deep Learning-based Identification of Cancer or Normal Tissue using Gene Expression Data. In: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM): IEEE; 2018. p. 1748–52.
9. Lyu B, Haque A. Deep learning based tumor type classification using gene expression data. In: Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics: ACM; 2018. p. 89–96.
10. Li Y, Kang K, Krahn JM, Croutwater N, Lee K, Umbach DM, Li L. A comprehensive genomic pan-cancer classification using the Cancer genome atlas gene expression data. BMC Genomics. 2017;18(1):508.
11. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, vol. 2017; 2017. p. 618–26.
12. Sun K, Wang J, Wang H, Sun H. GeneCT: a generalizable cancerous status and tissue origin classifier for pan-cancer biopsies. Bioinformatics. 2018; 34(23):4129–30.