Computational and Structural Biotechnology Journal 19 (2021) 835–846

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Genomic pan-cancer classification using image-based deep learning

Taoyu Ye, Sen Li, Yang Zhang *

Harbin Institute of Technology (Shenzhen), Shenzhen, Guangdong, 518055, China

ARTICLE INFO

Article history:
Received 26 October 2020
Received in revised form 5 January 2021
Accepted 8 January 2021
Available online 15 January 2021

Keywords:
Pan-cancer classification
Genetic mutation map
Image-based deep learning
Guided Grad-CAM visualization
Tumor-type-specific genes
Pathway analysis

ABSTRACT

Accurate cancer type classification based on genetic mutation can significantly facilitate cancer-related diagnosis. However, existing methods usually use feature selection combined with simple classifiers to quantify key mutated genes, resulting in poor classification performance. To circumvent this problem, a novel image-based deep learning strategy is employed to distinguish different types of cancer. Unlike conventional methods, we first convert gene mutation data containing single nucleotide polymorphisms, insertions and deletions into a genetic mutation map, and then apply the deep learning networks to classify different cancer types based on the mutation map. We outline these methods and present results obtained in training VGG-16, Inception-v3, ResNet-50 and Inception-ResNet-v2 neural networks to classify 36 types of cancer from 9047 patient samples. Our approach achieves overall higher accuracy (over 95%) compared with other widely adopted classification methods. Furthermore, we demonstrate the application of a Guided Grad-CAM visualization to generate heatmaps and identify the top-ranked tumor-type-specific genes and pathways. Experimental results on prostate and breast cancer demonstrate our method can be applied to various types of cancer. Powered by the deep learning, this approach can potentially provide a new solution for pan-cancer classification and cancer driver gene discovery. The source code and datasets supporting the study is available at https://github.com/yetaoyu/Genomic-pan-cancer-classification.

1. Introduction

Cancer is considered as the deadly genetic diseases, characterized by abnormal cell growths [1,2]. Globally, more than 18 million new cancer cases are diagnosed resulting to 9.6 million deaths in 2018 [3]. Genetic mutations have been shown to be associated with different types of cancer [4–6]. Cancer classification based on genetic mutations can be readily achieved through increased usage of high-throughput sequencing techniques. A large amount of mutation data has been generated and publicly released. Among them, The Cancer Genome Atlas (TCGA) is a cohort cataloguing genetic mutations data for more than 30 types of cancers from more than 10,000 patients [7]. TCGA contains various genetic mutations data, including single-nucleotide polymorphism (SNP), small insertions or deletions (INDEL), copy number variations (CNV), etc. By handling the massive amount of data, researchers now are able to design new analytical methods for accurate cancer classification and detection based on gene alteration. However, accurate and reliable cancer classification is particularly challenging as a result of the complexity and scale of the data. Considering the sequencing covers more than thousands of genes, but most of genes did not contain informative mutations thus making classification difficult by analyzing all those genes [8,9]. In order to avoid the mutation data being too sparse (even all zero), most analytical methods screen genes before classification [10,11]. These methods are simple and effective in some cases, but important features (genes) may be removed during the screening process.

Recent advances in deep learning underpin a collection of algorithms with an impressive ability to analyze molecular data without prior feature selection or human-directed training. Prior deep learning approaches usually work well for a specific type of cancer, such as brain cancer [12], gliomas [13], acute myeloid leukemia [14], breast cancer [15,16], soft tissue sarcomas [17] and lung cancer [18]. Given the complexity of pan-cancer data, directly using those mentioned approaches might not be appropriate for multiple types of cancer. Recently, some works are starting to consider the importance of genetic mutations in multiple types of cancer classification. By analyzing more than 8000 samples' genetic mutations profiles from 12 cancer types obtained from the TCGA, Sun et al. [19] reported a novel method, Genome Deep Learning (GDL), for

* Corresponding author.
E-mail addresses: zhangyang07@hit.edu.cn, yang.zhang2020@hotmail.com (Y. Zhang).

https://doi.org/10.1016/j.csbj.2021.01.010

cancer subtyping. However, more than 12 specific models were constructed. Limited by the number of models, this approach will be insufficient and unconfident in analysis of more types, and larger cancer mutation data. Yuan et al. [20] described DeepGene, an advanced Deep Neural Network (DNN) based cancer type classifier. Experimental results on 12 selected types of cancer from TCGA demonstrated improved classification performance compared with classifiers of Support Vector Machine (SVM), k-Nearest Neighbors (KNN) and Naïve Bayes (NB). However, the DNN classifier only has the optimal accuracy of 65.5%, which will prevent its development as an accurate cancer classifier.

In addition, most of these studies usually used only one type of genetic mutation data as input for cancer classification, which limits the performance of the classifier. For instance, Yuan et al. proposed DeepGene on somatic point mutation data for cancer classification [20]. AlShibli et al. [21] proposed three deep learning techniques to classify six cancer types based on CNV data. Although these methods are effective, the characteristic information is still not comprehensive enough. As far as we known, there is no existing work specifically designed to combined multiple types of mutation data.

As a result, a general algorithm for easy and reliable cancer classification based on multiple types of genetic mutation data is still missing. Previous works tend to use a variety of modeling methods, sometimes combine them together. In such a context merely adopting deep learning approaches developed within other setting might not be appropriate in pan-cancer classification based on different gene mutation data. Given these challenges, a new and simple approach is necessary.

Motivated by works of deep learning in image analysis, we describe a novel image-based deep learning strategy for cancer classification and mutated gene discovery. The proposed strategy is consisting of three main steps: construction of genetic mutation map, classification using deep Convolutional Neural Networks (CNN) and identify cancer driver genes by Guided Grad-CAM (a combination of Guided backpropagation and Gradient-weighted Class Activation Mapping) visualization [22]. This novel strategy makes the following research contributions:

(1) A genetic mutation map was constructed for each cancer patient, documenting the gene alternations condition including single-nucleotide polymorphism (SNP), insertion (INS) and deletion (DEL) with chromosome position information. Prior knowledge on the mutated genes selected is not necessary, avoiding bias caused by hand-picking. The correlation between mutated genes and cancer types can be built without gene prescreening.
(2) Genetic mutation map and popular deep neural networks, which used in combination, produce a high accuracy in pan-cancer classification. Compared with other widely used classification methods (such as SVM and KNN), our test classifiers, including VGG-16 [23], Inception-v3 [24], ResNet-50 [25] and Inception-ResNet-v2 [26], can effectively extract deep features from complex genetic mutation data, and significantly improve the classification accuracy.
(3) The application of Guided Grad-CAM visualization to generate heatmaps were utilized to identify tumor type-specific genes and pathways.
(4) The systematical examination of gene mutations in 36 types of cancer from 9,047 patient samples demonstrates the advancement of our method, allowing a deeper understanding of the mutation landscape of cancer. The constructed genetic mutation map dataset was publicly released at https://github.com/yetaoyu/Genomic-pan-cancer-classification/tree/master/DNN-models/dataset.

2. Materials and methods

2.1. Cancer types and samples statistics

The genetic mutation data from various types of cancer in TCGA are collected from the Firebrowse portal (http://firebrowse.org/). The dataset is assembled by selecting the genes across all samples for 36 cancer types that contain mutations. As shown in Supplementary Fig. S1, the upper line chart represents the number of mutation genes from each type of cancer, and the lower bar chart represents the total number of mutation conditions from those genes, including SNP, INS and DEL. As shown in the horizontal axis, the sample number of each tumor type ranges from Cholangiocarcinoma (CHOL, n = 35) to Breast invasive carcinoma (BRCA, n = 982). From 9,047 TCGA samples with 23,231 mutation genes, we demonstrate the general applicability of our image-based deep learning method on 36 types of cancers.

2.2. Mutation map construction

To construct the mutation landscape of cancer, we create the genetic mutation map. Assuming that the size of the mutation map is $N \times N$, all the mutation genes from 36 types of cancers are collected, grouped and located to the matrix map according to their positions on the chromosomes.

Firstly, mutated genes from each type of cancer are sorted according to their positions on chromosome (chromosomes 1–22, X and Y). For cancer j, the list of mutated genes on chromosome i is $\boldsymbol{r_{ij}}$, where $0 \leqslant i \leqslant 23$, $0 \leqslant j \leqslant 35$ in this paper. Then the mutated genes in the same chromosome from different types of cancer are grouped according to their positions. For chromosome i, the length of mutated gene set $R_{i\cdot} = r_{i0} \cup r_{i1} \cup \cdots \cup r_{i35}$ collected from different cancers is L_i. Therefore, the number of columns occupied by the genes on chromosome i in the mutation map is $k_i \times 3$, where $k_i = \begin{cases} \lfloor L_i/N \rfloor + 1, & if \quad L_i\%N \neq 0 \\ L_i/N, & if \quad L_i\%N = 0 \end{cases}$. To be specifically, each gene occupies three pixels in the same row of the mutation map, where each pixel point represents the mutation condition of the gene. These three pixels are colored with blue, green or red to represent SNP, INS or DEL respectively. Genes on all chromosomes occupy K columns, where

$$K = \sum_{i=0}^{23} 3k_i = \sum_{i=0}^{23} 3 \times \begin{cases} \lfloor L_i/N \rfloor + 1, & if\ L_i\%N \neq 0 \\ L_i/N, & if\ L_i\%N = 0 \end{cases} \quad and\ K \leqslant N$$

According to the above description, we can choose an appropriate N value. In our experimental data, the value of N is 310. Next, a collection of mutation genes are arranged and aligned vertically in the matrix map, according to their positions and orders on the chromosomes 1 to 22, X, and Y, thereby forming a $N \times N$ genetic mutation map for each tumor sample (Fig. 1A). Each chromosome occupies $k_i \times 3$ columns in the genetic map, containing $p(p \leqslant k_i \times N)$ genes and the extra pixels in the image are set to zeros. Finally, we output the genetic mutation maps for all of 9,047 patient samples from 36 types of cancer. All the mutation maps are normalized by the maximum value over RGB channels.

2.3. Without mutation map construction

Given M patient samples, the input to machine learning models without mutation map construction is composed of the class labels, i.e., the tumor type for the M individuals, and then the counts for mutation of all genes from the M patient samples are fed into a classifiers as described below. Let Q be the number of genes for which SNP data is available. Let $\boldsymbol{C} \in \{0 \ldots (t-1)\}^M$ be the vector containing the class labels where t is the number of cancer types, and let $\boldsymbol{G} \in \{0 \ldots s\}^{M \times Q}, s \in \mathbb{N}$ be the matrix repre-

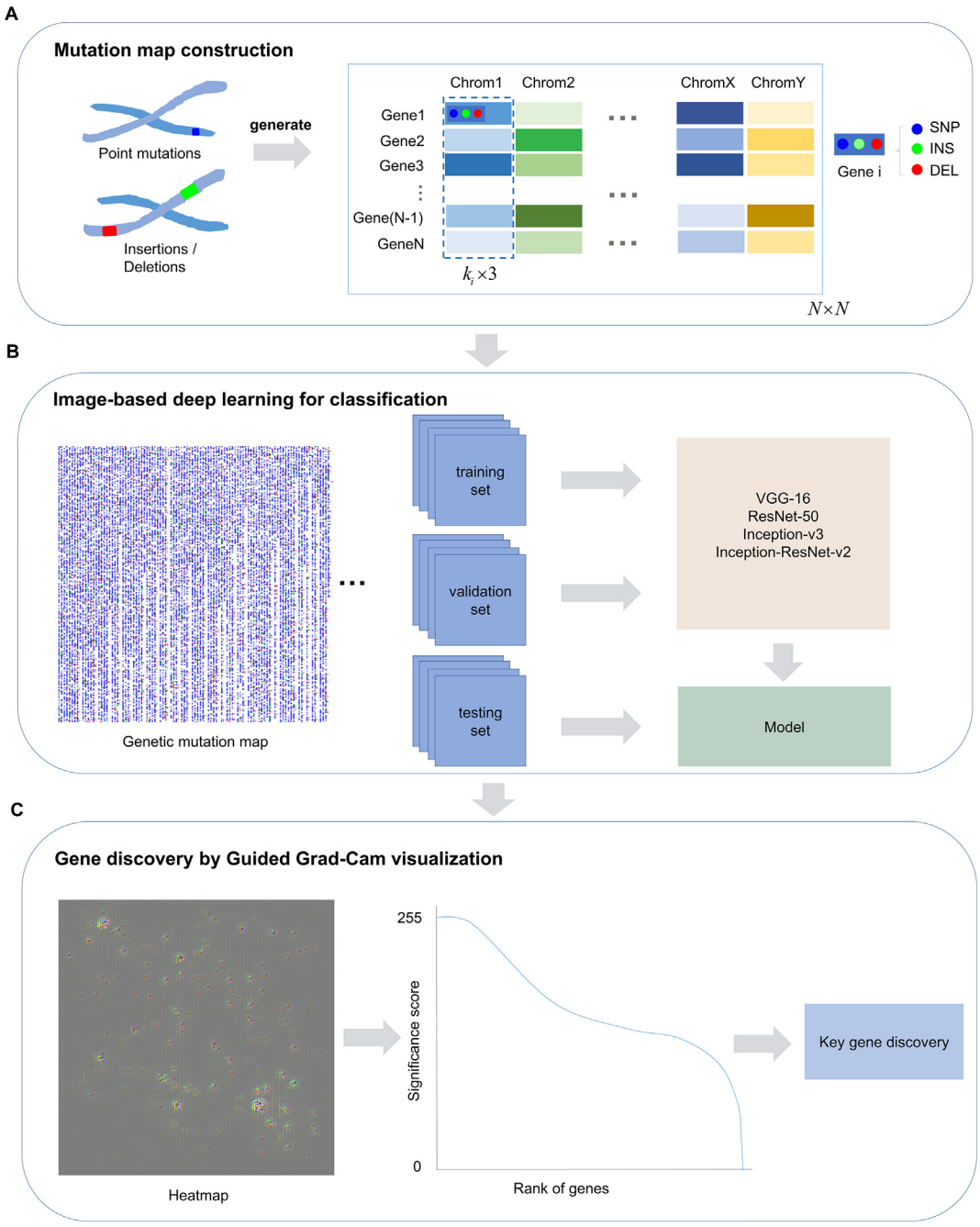

Fig. 1. Schematic representation of image-based deep learning for genomic pan-cancer classification. (A) The protocol of genetic mutation map construction. The gene mutation conditions including single-nucleotide polymorphism (SNP), insertion (INS) and deletion (DEL) with chromosome position information are transformed into the genetic mutation map. Each chromosome occupies $k_i \times 3$ columns in the genetic map, containing $p(p \leqslant k_i \times N)$ genes. Each gene occupies three pixels in the same row of the mutation map, where each pixel represents the mutation condition of the gene, colored blue, green, or red according to their labels to SNP, INS and DEL. Those pixel points are arranged and aligned vertically in the mutation map, according to their positions on the chromosomes, there forming a $N \times N$ matrix map for each patient, referred as the genetic mutation map. **(B)** Workflow of establishing the image-based deep learning models. All patient samples are transformed into the mutation maps and then divided into training, validation, and testing sets, respectively. The images of mutation maps are fed into different deep learning architectures for training and testing on pan-cancer classification. **(C)** Guided Grad-CAM are employed to generated heatmaps for the identification of top distinct candidate genes that help the pan-cancer classification. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

senting the number of mutations observed in each gene (i.e., $\boldsymbol{G}(m,q) = s$ if gene q has s mutations in sample m).

2.4. Machine learning methods

Logistic Regression (LR) is a supervised learning classification algorithm, which transforms its output using the logistic sigmoid function to predict the probability of two or more discrete classes [27].

The Bayesian classification represents a supervised learning method based on Bayesian Theorem, which combines knowledge of the distributions of feature vectors and the prior probabilities of the classes [28]. Naive Bayes assumes conditional independence and uses the method of maximum likelihood in parameter estimation. Gaussian Naive Bayes (GNB) is a variant of Naive Bayes that follows Gaussian normal distribution.

KNN is a basic, supervised machine learning algorithm that is widely used for classification and pattern recognition [29]. KNN

classifier calculates the similarity between a new data point to training data points based on distance measurement. In this study, the nearness of points is determined by Euclidean distance formula [30].

SVM, a supervised learning algorithm, is developed to classify both linear and nonlinear data, which transforms the original data into a higher dimension, from where it can find an optimal separating hyperplane between the classes using essential training tuples called support vectors [31,32]. In this study, we use Radial Basis Function (RBF) kernel which transforms the data points to higher-dimensional space.

Random Forest (RF), one of the most used supervised classification algorithms that constructs multiple decision trees, using random sampling to select subset of training data and variables [33]. RF is an ensemble of many decision trees and uses the Gini index to measure the impurity of a node in deciding its splitting.

Gradient Boosting Decision Tree (GBDT) is also an ensemble learning by combining multiple decision trees for classification. GBDT is sequential ensemble learning technique where the performance of the model improves over iterations. This method constructs the model in a stage-wise fashion, by combining a set of weak learners into a single strong learner through iterative methods [34].

In this paper, the parameter settings used in LR, GNB, KNN, SVM with RBF kernel and RF are the same as in the [35]. The GBDT sets the maximum number of iterations of the weak learner n_estimators = 200. All the classifiers were available from the Python package scikit-learn.

Most feature selection methods tend to perform effective screening on genes and sample data in advance, and then uses those features as input to the classifiers, however, machine learning methods with mutation map conversion uses the whole data set as input. Therefore, in the comparative experiments on 36 types of cancer data with or without mutation map transformation, we did not perform feature selection for the classifiers without mutation map transformation, ensuring the consistency of input dataset.

2.5. Deep learning methods

The VGG-16 network is composed of 16 convolutional layers and has a small receptive field of 3×3. In one of the configurations, it also utilizes 1×1 convolution filters. It has five max-pooling layers of size 2×2. There are 3 fully connected layers after the last max-pooling layer. It uses the softmax layer as the final layer. All hidden layers are equipped with the Rectified Linear Units (ReLU) activation. The Inception-v3 network introduces the asymmetric decomposition of the convolution kernel, which replaces any $n \times n$ convolution by a $1 \times n$ convolution followed by a $n \times 1$ convolution. The model is a CNN with 48 layers and consists of several Inception modules. As to ResNet-50 model, a deep residual network with a depth of 50 layers and residual connection is the most significant idea in ResNet design [30]. A direct connection branch is added between some weight layers, that is, the input features are directly connected to the output to form a complete residual connection module.

Suppose input images are $\boldsymbol{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_k, \cdots, \boldsymbol{x}_M\} | (1 \leqslant k \leqslant M)$ with their class labels $\boldsymbol{Y} = \{\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_k, \cdots, \boldsymbol{Y}_M\} | (1 \leqslant k \leqslant M)$, where M is the number of images. The convolutional neural network mentioned above are used to extract discriminative features of $\boldsymbol{X}$ iteratively, obtaining a satisfactory classification result of features. Denote there are L layers in the convolutional neural network. For image $\boldsymbol{x}_k$, the output of the first layer is $\boldsymbol{h}_k^1 = \sigma\left(\boldsymbol{w}^1 \boldsymbol{x}_k + \boldsymbol{b}^1\right)$, where w^1, b^1 are the convolutional weight and bias in the first layer respectively, and σ is the activation function. So $\boldsymbol{h}_k^l = \sigma\left(\boldsymbol{w}^l \boldsymbol{h}_k^{l-1} + \boldsymbol{b}^l\right)$ is the output of the $l-th$ layer $(1 < l < L)$ for $\boldsymbol{x}_k$. The feature representation of $\boldsymbol{x}_k$ from the last layer, is $\boldsymbol{h}_k^L = \sigma\left(\boldsymbol{w}^L \boldsymbol{h}_k^{L-1} + \boldsymbol{b}^L\right)$, which is fed into a softmax loss function,

$$\mathscr{L} = -\sum_{k=1}^{M} \mathrm{y_k} \log \mathbf{h}_k^{\mathrm{L}}$$

From this loss function, the last layer of the network can output the predicted category given an image. As for the network difference, Inception-Resnet-v2 is 164 layers deep and the basic structure is similar to Inception, but the residual connection technology is added to the original Inception module to further improve the network convergence speed and accuracy.

2.6. Image-based deep learning for pan-cancer classification

After constructing mutation map, we randomly partitioned all 9,047 mutation maps into a training, validation and test datasets (Supplementary Table S1). Four popular image recognition deep learning models, including VGG-16, Inception-v3, ResNet-50 and Inception-ResNet-v2, are compared. These models are directly called by keras applications package and have several hyperparameters, including learning rate, optimizer, epochs, batch size, decay factor, and others. These hyperparameters determined by using ten-fold cross-validation were summarized in Supplementary Table S2. In all CNN models, parameters are trained using the training set and tuned using the validation set. The training set weights are saved when the loss for the validation dataset is convergence. In the evaluation using the test dataset, the weights are loaded. The CNN classifier is subsequently used to predict the class probability of the samples in the testing set. The predicted and true class memberships are then compared to calculate the testing set prediction accuracy, precision, recall and F1-score. $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$, $Precision = \frac{TP}{TP+FP}$, $Recall = \frac{TP}{TP+FN}$, $F1-score = \frac{2 \times Precision \times Recall}{Precision+Recall}$. Where TP, TN, FP, FN are true positive, true negative, false positive and false negative. When calculating the precision, recall and F1-score values of each experiment, considering the imbalance of the sample, we set average="weighted". All the evaluation metrics were available from the Python package scikit-learn metrics.

The prediction results may vary depending on which samples are assigned to the training set. Given the large size of the cancer genetic map dataset and the high computing resource demanding of the algorithm, we repeated the above procedure three times, each with an independent training/validation/testing partition to avoid idiosyncrasies from use of a single random assignment. By multiplying experiments with maximum epochs as 1,000, we found that the model has converged when the number of epochs is less than 100 and has remained relatively stable after converged. Therefore, we finally use 100 as the number of training epochs for the initialization of models. All results presented in the manuscript are based on samples from testing sets that are not involved in the training process.

To compare the performance of the four models, we used the same hyperparameter settings. In each model, we used random initialization weights. Then, the weights of the whole model were trained together. Stochastic Gradient Descent (SGD) method was employed to tune the parameters of the models. The learning rate was set to 1e−3, and the batch size was set to 16. To improve the performance of models, we also utilized the ReduceLRonPlateau schedule with patience of 3 epochs, decay factor of 0.5 to dynamic decrease the learning rate. The learning rate will be decayed to 1e−5 along with the training. In addition, we adopted an early stopping strategy with patience of 5 epochs to effectively prevent

overfitting. The training will stop when the minimum validation loss is not improving in 5 rounds or epochs are up to 100. Although Adam optimizer is popular, SGD optimizer is more suitable for fine-tuning parameters. As shown in the Supplementary Table S3 and Supplementary Figs. S2–S4, all four deep learning models have achieved excellent results when using SGD but not Adam.

2.7. Method for heatmaps generation

Visualization techniques Guided Backpropagation [36], Gradient-weighted Class Activation Mapping (Grad-CAM) or Guided Grad-CAM have been used to better understand decisions made by deep convolutional neural networks during the pan-cancer classification. Back-propagation is a neural network algorithm employing gradient descent for classification. Rules are extracted from trained neural networks to help improve the interpretability and visualization of the learned network [37].

Grad-CAM is also a visual interpretation of CNNs based on gradients of targets. Given a specific category and layer, Grad-CAM can perform weighted summation on the feature maps in the convolutional network to obtain the channel weights of the layer and can further produce a localization map highlighting important regions in the image [22].

The final heatmaps are generated by Guided Grad-CAM, a combination of Guided Backpropagation and Grad-CAM. We input all the testing samples into the trained CNNs. For each patient sample, we record the activation map of the last convolutional layer during the forward passing process. After that, we further record the label-specific gradient of each neuron in the last convolutional layer through the back-propagation process. These gradients represent the contribution of neurons to the classification results [38]. Using a weighted sum of activation maps to calculate Grad-CAM, and then multiply the gradient of the input layer to generate the Guided Grad-CAM heatmap for each sample. Finally, we average all the testing heatmaps from the same category and obtain the heatmap of each category after the Min-Max normalization. The intensity of each pixel represents the significance score of the corresponding gene to the pan-cancer classification.

2.8. Validation of top genes

Top genes are selected according to the ranking of significance scores in the heatmaps. We apply functional analysis on these top genes to further prove that the genes are tumor-type specific. Top genes within 1,000 from the prostate and breast cancer are selected and validated to find their relations to the corresponding tumor.

To obtain a functional representation of the lists of potential mutation genes identified by heatmap, we perform Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways enrichment analyses using the online database and tool DAVID (version 6.8, https://david.ncifcrf.gov).

2.9. PC and software used for calculation

For all calculations, a custom-mode PC with a CPU (Core i7-8700 Intel Xeon CPU E5-2640 v4 @ 2.40 GHz) and 4 GPU (Tesla K80, NVIDIA GK210GL) are used. The installed OS is Ubuntu 16.04 LTS. We use TensorFlow- version 1.12.0 (https://github.com/tensorflow/tensorflow/releases/tag/v1.12.0) and Keras version 2.2.5 (https://github.com/keras-team/keras/releases/tag/2.2.5) to build the models.

3. Results

3.1. Overall design of image-based deep learning

Our newly developed model is an image-based deep learning approach for pan-cancer classification and gene mutation discovery, consisting of three main steps:

1) Transformation of gene mutation data into genetic mutation map (Fig. 1A). For each patient sample, a genetic mutation map is constructed, documenting the gene alternations condition including SNP, INS and DEL with chromosome position information.
2) Genomic pan-cancer classification using image-based deep learning. Taking the obtained genetic mutation map as input, we next train a classifier using deep learning networks. By comparing trainable multilayer convolutional neural networks, including popular VGG-16, Inception-v3, ResNet-50, and Inception-ResNet-v2 networks, deep learning networks takes the advantage of genetic mutation map, which are more comprehensive and informative (Fig. 1B).
3) To identify the key genes that aiding cancer classification in the genetic mutation map, we then employ Guided Grad-CAM visualization to generate the heatmap. By inspecting the produced heatmaps, we obtain localized discriminative molecular patterns extracted from the original maps that help the CNN classification. The results imply that a visual discriminative pattern or pixel could be used in order to discover what sort of genes that most strongly associated with one particular type of cancer (Fig. 1C).

3.2. The importance of mutation map construction

To evaluate the importance of mutation map, we first test several widely used machine learning methods on 36 types of cancer data with or without mutation map transformation. As shown in Table 1, all classifiers with mutation map construction show better performance in terms of accuracy and F1-score. This might be the reason that data without mutation map transformation are essentially row vectors with limited information, compared to ones with mutation map conversion are two-dimensional vectors containing more mutation information. In addition, the feature matrix is more sparseness due to feature selection is not performed in advance, so it is difficult to extract meaningful features.

Although mutation map construction results in relatively significantly improved in performance with LR and GNB method, it can more dramatically enhances the classification performance of KNN, SVM with RBF kernel, RF and GBDT algorithm. The methods outlined above achieve the best classification accuracy of 93.1% in SVM-based image analysis. Improved classification results on SVM-based method, and other various methods suggest an important role of mutation map construction. However, those traditional classifiers cannot identify genes that most strongly associate with one particular type of cancer. Therefore, a new approach is necessary for pan-cancer classification and also gene mutation discovery with a high accuracy. Deep neural networks have recently attracted great attention, and might potentially provide a new solution for this challenge.

3.3. Image-based deep learning model for pan-cancer classification

Although the traditional machine learning methods with mutation map construction can achieve high accuracy, they cannot further identify candidate genes that help the classification. Widely used deep learning networks including VGG-16, Inception-v3,

Table 1
Evaluation of pan-cancer classification performance by different models. Average testing accuracy, precision, recall and F1-score of several machine learning and deep learning methods on 36 types of cancer with or without mutation map transformation are calculated from three independent experiments.

Method	Mutation map Construction	Accuracy(%) (±SD(%))	Precision(%) (±SD(%))	Recall(%) (±SD(%))	F1-score(%) (±SD(%))
Machine learning methods					
LR	w/o	37.18 ± 0.818	39.56 ± 0.331	37.18 ± 0.818	37.17 ± 0.414
	w/	59.78 ± 0.563	66.18 ± 0.268	59.78 ± 0.563	61.23 ± 0.424
GNB	w/o	12.81 ± 0.000	13.98 ± 0.002	12.81 ± 0.000	10.66 ± 0.001
	w/	32.43 ± 0.002	30.99 ± 0.001	32.43 ± 0.002	29.85 ± 0.000
KNN	w/o	5.63 ± 0.279	10.93 ± 0.224	5.63 ± 0.279	3.03 ± 0.165
	w/	66.85 ± 0.002	69.85 ± 0.003	66.85 ± 0.002	66.72 ± 0.002
SVM	w/o	34.03 ± 0.139	40.27 ± 0.297	34.03 ± 0.139	34.67 ± 0.161
	w/	93.11 ± 0.003	93.37 ± 0.002	93.11 ± 0.003	92.98 ± 0.001
RF	w/o	30.18 ± 0.158	38.87 ± 0.564	30.18 ± 0.158	28.38 ± 0.213
	w/	92.69 ± 0.049	93.14 ± 0.028	92.69 ± 0.049	92.52 ± 0.040
GBDT	w/o	32.88 ± 0.169	37.91 ± 0.609	32.88 ± 0.169	33.08 ± 0.249
	w/	86.66 ± 0.123	87.12 ± 0.039	86.66 ± 0.123	86.48 ± 0.096
Deep learning methods					
VGG-16	w/	99.71 ± 0.075	99.72 ± 0.075	99.71 ± 0.075	99.71 ± 0.075
Inception-v3	w/	95.33 ± 1.494	95.52 ± 1.445	95.33 ± 1.494	95.12 ± 1.582
ResNet-50	w/	99.63 ± 0.038	99.63 ± 0.036	99.63 ± 0.038	99.63 ± 0.038
Inception-ResNet-V2	w/	99.12 ± 0.113	99.22 ± 0.057	99.12 ± 0.113	99.14 ± 0.097

'w/': with mutation map transformation. 'w/o' : without mutation map transformation. SD = Standard deviation. LR : Logistic Regression. GNB : Gaussian Naive Bayes. KNN : k-Nearest Neighbors. SVM : Support Vector Machine. RF : Random Forest. GBDT : Gradient Boosting Decision Tree.

ResNet-50 and Inception-ResNet-v2 are tested with the genetic mutation maps for pan-cancer classification (Fig. 1B). Genetic mutation maps of 36 types of cancer patients are constructed and further subgrouped into training, validation, and testing dataset (see Supplementary Table S1). Accuracy and cross-entropy loss in the training and validation datasets are plotted against the training step during the training of different deep learning networks (Supplementary Fig. S2). The training is stopped when the loss does not improve in 5 epochs or epochs are up to 100 (iterations through the entire dataset). In comparison with machine learning methods, all deep learning algorithms achieve higher classification accuracies, ranging from 95.33 to 99.71 (Table 1). Area Under the Curve (AUC) in the Receiver Operating Characteristic (ROC) curves achieve the highest accuracies of 1.0 for four deep learning models (Supplementary Fig. S5).

To further illustrate the accuracy of these predictions, confusion matrixes are plotted, suggesting that almost all samples of cancer types could be correctly classified (Fig. 2). What more interesting is that high classification accuracies were achieved with cancer types of smaller sample sizes, such as Cholangiocarcinoma (CHOL, n = 35), Lymphoid neoplasm diffuse large B-cell lymphoma (DLBC, n = 48), Uterine carcinosarcoma (UCS, n = 57), Kidney chromophobe (KICH, n = 66), Rectum adenocarcinoma (READ, n = 69). A close examination of the misclassified samples suggests the main reason might be similarity in the tumor types. As shown in Fig. 2, Colon adenocarcinoma (COAD) and Colorectal adenocarcinoma (COADREAD) are misclassified, both of which originated from colorectal tissues. In addition, the samples of Glioma (GBMLGG) are mis-assigned to Brain lower grade glioma (LGG), as both types of samples are brain tumors. Kidney renal papillary cell carcinoma (KIRP) samples are misclassified as the Pan-kidney cohort (KIPAN), which including KICH, Kidney renal clear cell carcinoma (KIRC) and KIRP tumors.

To examine the extent to which the image-based deep learning model is capable of distinguishing different types of cancer, we visualize the internal features revealed by the networks, using t-distributed stochastic neighbor embedding (t-SNE) [39] (Supplementary Fig. S6). The output features before the last classification features to the layer are compressed into two dimensions by t-SNE. A total of 1,250 test data maps are compressed and plotted in two-dimensional space. 36 types of cancer are clearly separated into different clusters in different image-based deep learning models, further suggests that the features learned through training can be used for pan-cancer classification. Cluster points are closer between similar types of cancer in the middle of the plot, such as COAD to COADREAD, GBM to GBMLGG, which are consistent with the results from confusion matrixes.

3.4. Gene mutation discovery

Neural networks have often been thought of as black boxes due to the difficult in understanding what and how they learn. To understand how those deep learning algorithms classify the mutation maps, a number of visual explanations are produced to highlight the discriminative region where the deep learning networks focused upon the classification. Grad-CAM, a class-discriminative localization technique, uses the gradient information flowing into the final convolutional layer of the networks to produce a coarse localization map highlighting the important regions in the image for pan-cancer classification by deep learning networks.

Using Prostate Adenocarcinoma (PRAD) and Breast Cancer (BRCA) samples as an example, heatmaps of deep learning algorithms VGG-16, Inception-v3, ResNet-50, and Inception-ResNet-v2 were generated. As shown in Fig. 3(b,d,f,h) and Supplementary Fig. S7(b,d,f,h), Grad-CAM can easily localize the regions for corresponding cancer classification; however, it is difficult to accurately localize the key discriminative pixels (genes) in the low-resolution heatmaps. While Grad-CAM visualizations are class-discriminative and localize relevant image regions well, they lack the ability to show fine-grained details. Guided Backpropagation are pixel-space gradient high-resolution visualization method, but are not class-discriminative [22].

Therefore, Guided Grad-CAM visualization is employed by combining both Guided Backpropagation and Grad-CAM visualizations via point-wise multiplication to create high-resolution and class-discriminative visualization. As shown in Fig. 3(c,e,g,i) and Supplementary Fig. S7(c,e,g,i), Guided Grad-CAM highlights fine-grained details in the image with high resolution compared with Grad-CAM in Fig. 3(b,d,f,h) and Supplementary Fig. S7(b,d,f,h). The heatmap could help to locate discriminative molecular patterns extracted from the original mutation maps that aiding the pan-cancer classifications. Top distinct candidate genes relative to the original pixel in the map can then be identified. By comparing the Guided Grad-CAM visualization generated by different deep

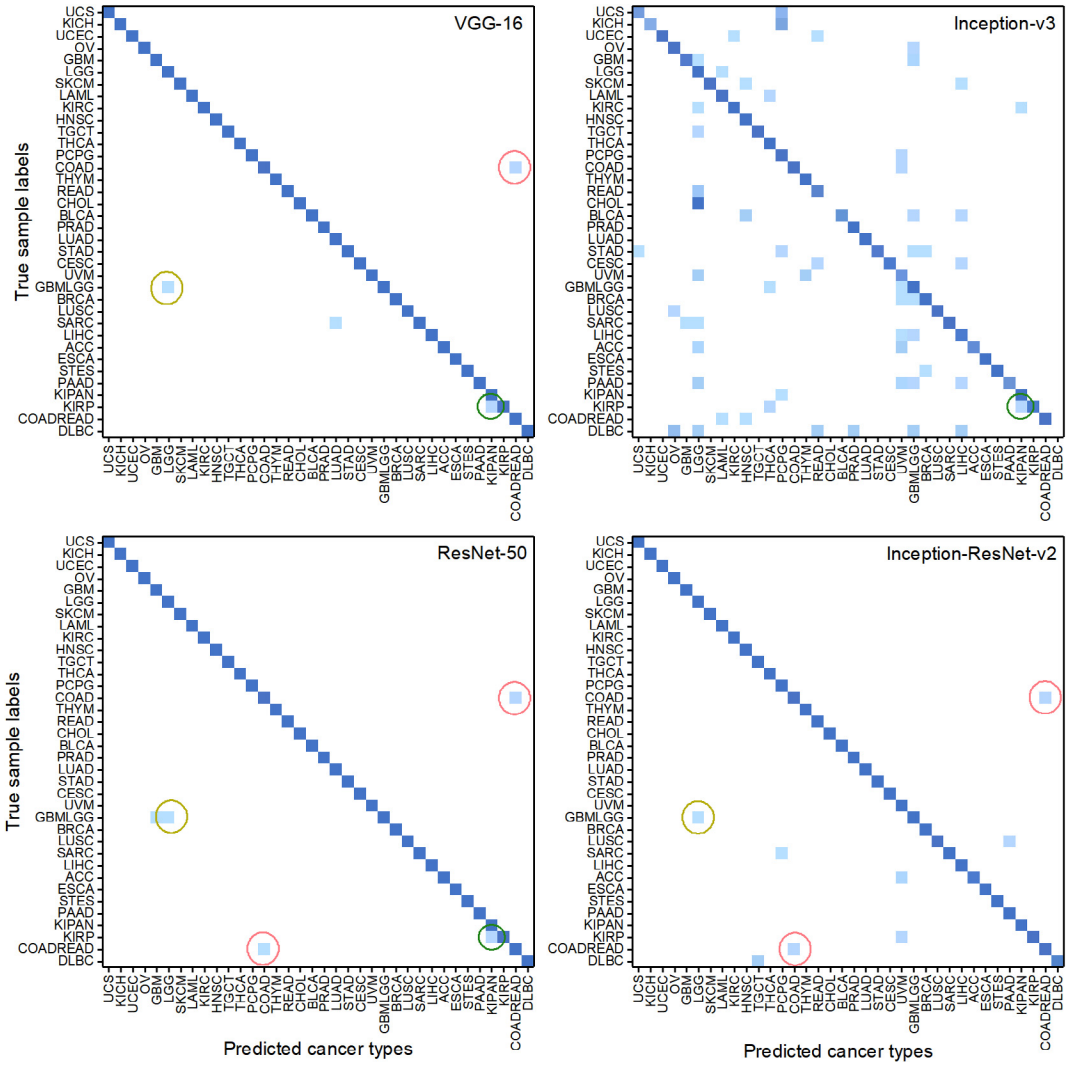

Fig. 2. The confusion matrix for pan-cancer classification. Rows represent the true classification of cancer types, and the columns represent the predicted cancer types. The diagonal cells correspond to cancer types are correctly classified. The off-diagonal cells correspond to incorrectly classified samples. Different colored circles indicate the similarity of misclassification between different models.

learning algorithms in a specific type of cancer, we find that there are similar discrimination visualization areas between different patient samples by the same model (Fig. 3 or Supplementary Fig. S7, rows). However, for the same sample, the molecular patterns recognized by different networks are quite different (Fig. 3 or Supplementary Figure S7, columns). There are some overlapping areas or patterns need to be identified.

3.5. Validation of top-ranked discriminative genes

The size of pixel values in the heatmap are used as an indicative of the importance of the corresponding genes for pan-cancer classification. Using prostate cancer and breast cancer as examples, the gene significance score (pixel value) curves of four algorithms are plotted (Fig. 4A and B). According to Fig. 4A and 4B, we find that when the significance score is greater than 150, the declining slope decline is more obvious. Therefore, the average heatmap of the PRAD (46 patient samples) and BRCA (136 patient samples) generated by four models are set to a threshold of 150 for fixed threshold segmentation, and average heatmap schematics Fig. 4C and 4D are obtained. In the Guided Grad-CAM heatmap, we find that the declining trends of the significance score curves of the four models are quite similar, first decrease sharply, and then become gradually smooth from around 1000th genes. The slopes of the intensity change in the first 1,000 genes are larger than the following several thousand genes, therefore we chose the top genes within 1,000 for further validation. The identified top ranked genes within first 1,000 genes are then uploaded to the online software DAVID database (https://david.ncifcrf.gov/) for KEGG pathway analysis. The DAVID database is a widely used gene enrichment and biological functional annotation database integrating comprehensive set of functional annotation tools for high-throughput gene function analysis [40]. The enriched pathways and corresponding cancer driver genes are identified. Literature searching confirms most of those pathways are oncogenic signaling pathways reported in prostate cancer (Fig. 5, Table 2) and breast cancer (Supplementary Fig. S8, Supplementary Table S4). In comparing the difference between different models, we overlap the KEGG pathways analyses based on top ranked genes inputs. As shown in Fig. 5 for prostate cancer, there is overlap in pathways between different models, mainly enriched in the PI3K-Akt (four models), focal adhesion (VGG-16, Inception-v3 and ResNet-50), extracellular matrix (ECM)-receptor interaction (Inception-v3 and ResNet-50), olfactory transduction (Inception-v3 and ResNet-50) signaling pathways, etc.

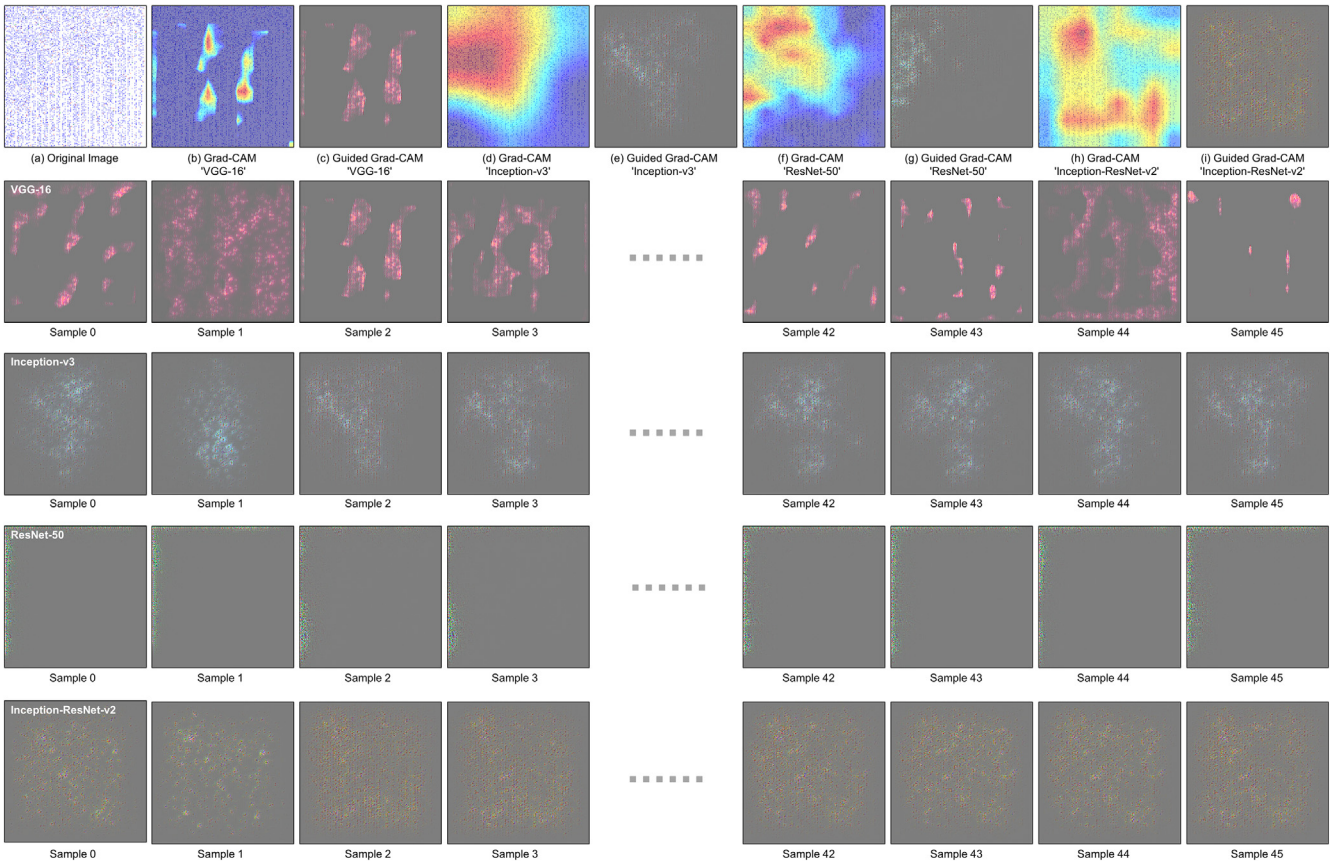

Fig. 3. Generation of heatmaps for deep learning networks visualization in prostate cancer. (a) Original genetic mutation map of a prostate cancer sample TCGA-HC-8264-01_PRAD. (b-c) Support for the prostate cancer classification according to different visualizations for VGG-16. (b) Grad-CAM: localizes class-discriminative regions. (c) Combining (b) and Guided Backpropagation gives Guided Grad-CAM, which gives high-resolution class-discriminative visualizations. (d-e) are different visualizations for Inception-v3. (f-g) are different visualizations for ResNet-50. (h-i) are different visualizations for Inception-ResNet-v2. Note that in (b,d,f,h), red regions correspond to high score for class. The rows represent the heatmaps representation of prostate cancer samples generated by the corresponding model. The columns represent the heatmaps representation of the same patient sample generated by different models. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Among them, the PI3K-Akt pathway is the best characterized somatic genetic mutation pathway. The well-known cancer pathways PI3K–Akt controls multiple cellular processes, and plays an important role in the occurrence and development of prostate cancer [41]. In prostate cancer, the frequency of PI3K pathway alteration rises substantially when mutations occurred in the INPP4B and PHLPP, the PIK3CA gene itself, and the PIK3CA regulatory subunits PIK3R1 and PIK3R3 [42]. Interestingly, our image-based deep learning approaches also identify different prostate cancer driver genes affecting the PI3K-Akt pathway, including PIK3R3, PIK3CD, PIK3C2B, PIK3AP1, PHLPP2, INPP5B, and many other genetic alterations (Table 2). Olfactory Receptors (ORs), belonging to the class of G-protein-coupled Receptors (GPCRs), plays an important role in tumor progression by inducing cell invasiveness through GPCRs activation of PI3K pathway in prostate cancer [43,44].

Further data analysis in breast cancer reveals that similar pathway alterations (PI3K-Akt and olfactory transduction pathways) occur in different type of tumor, but altered genes affecting those pathways are different (Supplementary Fig. S8, Supplementary Table S4). By comparing differences in pathways between those two different types of cancers, many other signaling pathways are also involved in the pathogenesis of breast cancer, such as Ras, MAPK, TGF-beta, AMPK signaling pathway, etc. The understanding of these additional pathways and altered genes may assist in the regulation mechanisms in breast cancer.

These results demonstrate different deep learning models can extract the distinct molecular patterns from heatmaps to correlate with the underlying biological characteristics of cancer subtypes. The image-based deep learning approaches when combined can draw the general landscape of cancers and identify hundreds of cancer driver genes. Although the roles of many of them need further clarification, several major oncogenetic signaling pathways seem to be altered.

4. Conclusions and discussions

Various computational approaches have been reported to classify different types of cancers, but few single methods that are designed for all cancer types. TCGA documents comprehensive, well-curated genomic data of over 10,000 patient samples across more than 30 types of cancers. Taking a classic machine learning approach to such dataset often requires to perform feature selection in advance, which is cumbersome.

In the present work, we describe a novel image-based deep learning approach for genomic pan-cancer classification. The main novelty of the paper is the proposal of constructing the genetic mutation map and then fed into the deep learning networks. Each pixel in the mutation map represents the mutation conditions of a gene, colored with blue, green, or red according to their labels to SNP, INS, or DEL. Those pixel points are arranged and aligned

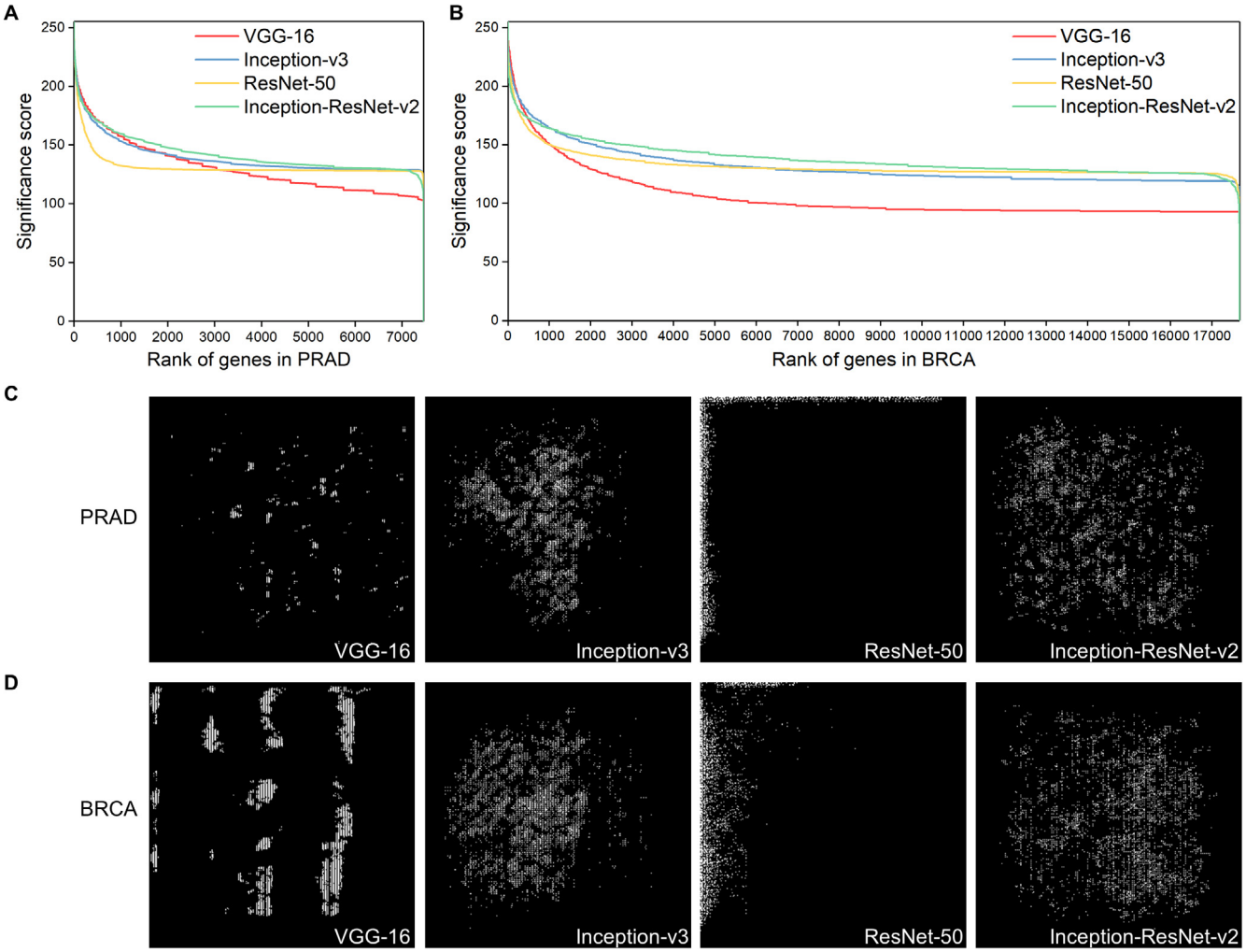

Fig. 4. The changes of gene significance scores for prostate adenocarcinoma (PRAD) **(A)** and breast cancer (BRCA) **(B)** with four deep learning models. The x-axis indicates the rank of genes and the y-axis denotes significance score of heatmap. The average heatmap schematics of the PRAD **(C)** and BRCA **(D)** are generated from 46 PRAD patient samples and 136 BRCA patient samples by four models.

vertically in the mutation map, according to their positions on the chromosomes. The genetic mutation data from 36 types of cancer in TCGA are evaluated to demonstrate the advancement of our method. Our approach achieves overall higher accuracy (over 95%) compared with other widely adopted classification methods such as LR, KNN, SVM with RBF kernel, RF and GBDT algorithms. With mutation map construction, all tested machine learning approaches exhibit improved classification accuracies, which serve as an outstanding tool for pan-cancer classification.

In contrast to those traditional machine learning methods, our image-based deep learning approach is able to discover a complete catalog of genes truly associated with cancer. The approach takes advantage of deep learning approach in image analysis for pan-cancer classification. We compare the performance of several popular deep learning frameworks, including VGG-16, Inception-v3, ResNet-50, and Inception-ResNet-v2. Genetic mutation map and deep neural networks, when used in combination, produce a high accuracy in classification, thereby illustrating the power and the generality of image-based deep learning approach. Moreover, different mutation conditions, including SNP, INS, and DEL, can be plotted in a single image and trained in a joint manner for pan-cancer classification. One major advantage of the use of genetic mutation map over mutation data is its ability to directly visualize deep learning networks. Different from the traditional approaches, prior feature selection is not necessary, avoiding bias caused by human-directed training. As a result, the proposed combinatorial approach will enable comprehensive genomic profiling in cancer. The systematical examination of mutation genes in large-scale cancer patients will potentially enable to prioritize cancer-causing genes, allowing a deeper understanding of the mutation landscape of cancer.

To extract the discriminative molecular patterns from the original maps that help the deep networks classifications, we utilize Guided Grad-CAM visualization. The networks output the precise pixel position of the potential key genes that could be associated with a type of cancer. This approach is more generally applicable and can be tested on any type of cancer. As a proof of concept, we have successfully applied the system to prostate cancer and breast cancer. The genetic mutation map is learned through four deep learning models to generate heatmaps. The top-ranked discriminative genes associated with prostate and breast cancer are determined by analyzing the heatmap generated by Guided Grad-CAM visualization.

Although different deep learning networks are able to taught themselves to accurately predict the cancer type, the general observation from heatmap is that networks do not necessarily have to have the same rules for classification. In terms of top-ranked discriminative gene discovery, we find those deep learning

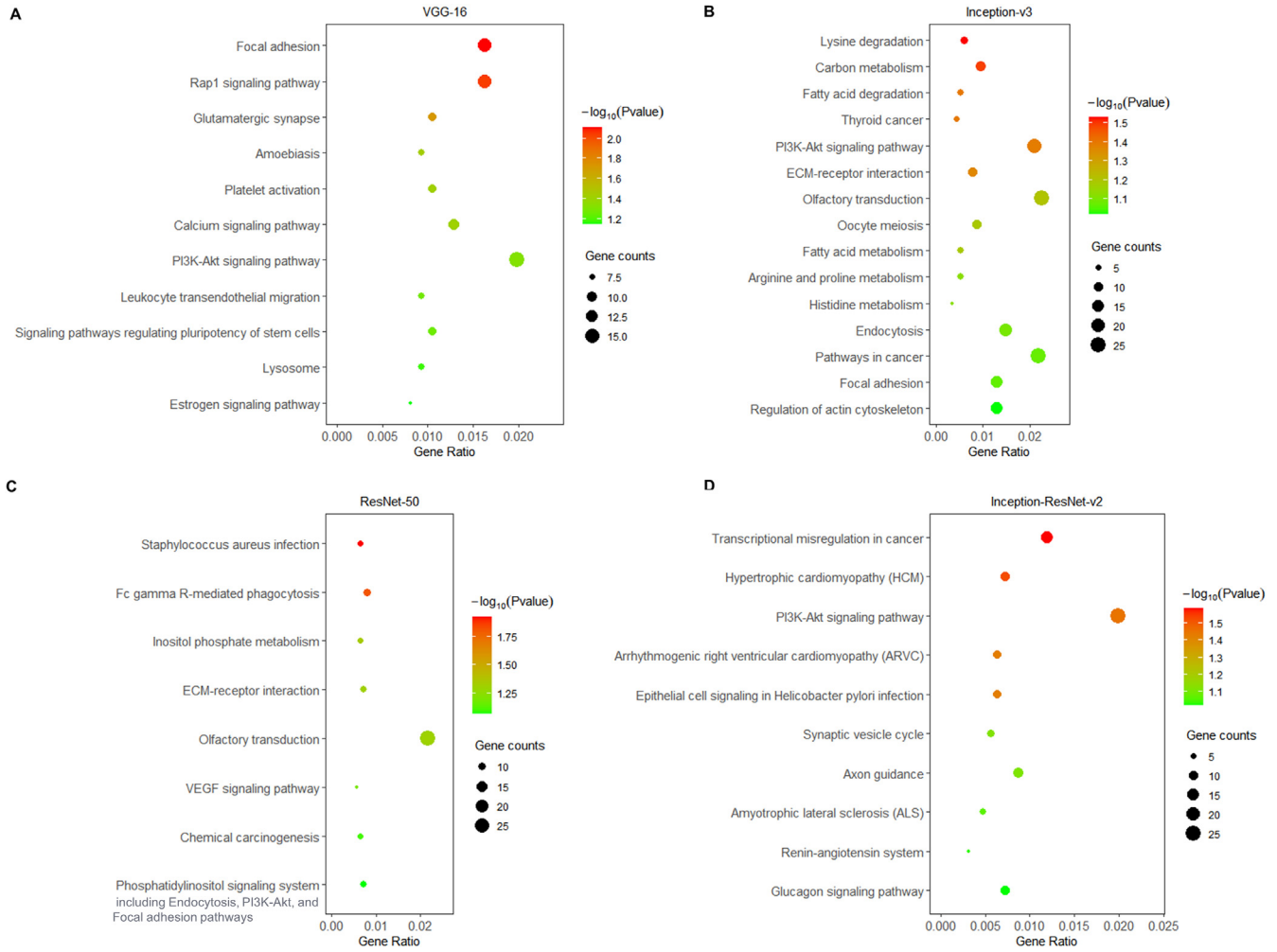

Fig. 5. KEGG pathway enrichment analysis of top-ranked prostate cancer driver genes in four deep learning models. (A) VGG-16 model. (B) Inception-v3 model. (C) ResNet-50 model. (D) Inception-ResNet-v2 model. The vertical ordinates are the terms of the KEGG pathways. Gene ratio is the proportion of the number of genes vs. the total number of genes in the same KEGG pathway. The color represents p-value. Gene counts represent the number of genes enriched in the pathway.

methods performance varies substantially across the examined type of cancer. The features used by different deep learning networks for classification might be different. It is unclear which patterns or learning rules are adapted by different deep networks. Changes in the networks structure, or in the learning processing, together with the stochastic nature of the optimization procedures, might also produce notably different results, making it extremely difficult to sift methods that significantly outperform others. Therefore, it is likely that some of methods are complementary to each other, and integration of such methods are able to identify a complete catalog of genes truly associated with cancer.

Herein, using prostate cancer and breast cancer as examples, we are able to identify significant enriched pathways of PI3K-Akt, olfactory transduction, and many other oncogenic pathways. Although some pathways and altered genes affecting those pathways have been previously reported to be associated with cancer, the proposed combinatorial approach reveals potentially novel pathways and cancer driver genes. The roles of these novel gene candidates in cancer need to be confirmed experimentally. Our results demonstrate that using the image-based deep learning approach can successfully identify biologically relevant and tumor-type-specific gene mutations.

However, the proposed approach has some limitations such as incapability of analyzing a small dataset due to the requirement of the size of the model input picture. In addition, the internal rules in our image-based deep learning models for pan-cancer classification and key genes discovery are not fully investigated. Labeled color and gene arrangement in the mutation map might be also important features used by the deep-learning model for classification. Guided Grad-CAM visualization can highlight pixels (genes) in the heatmaps are most strongly associated with one particular type of cancer, but the reason why those areas or pixels used by the deep learning models are not well understood. Therefore, to build an integrative model for accurate cancer profiling, further experiments understanding those factors need to be carried out.

In summary, we represent a generalized approach that can potentially be applied to a wide range of molecular data (e.g., gene expression, copy number variation, DNA methylation) for multiple types of disease classification. Our image-based deep learning approach can be a useful tool to assist pathologists in disease classification and the discovery of disease-causing molecules.

Funding

This work was supported by the Natural Science Foundation of Shenzhen City [grant number JCYJ20180306172131515].

CRediT authorship contribution statement

Taoyu Ye: Data curation, Formal analysis, Methodology, Project administration, Resources, Software, Validation, Visualization,

Table 2
KEGG pathway analysis of top-ranked prostate cancer driver genes in four deep learning models.

Model	KEGG terms	Ref	P-value	Genes
VGG-16	hsa04510:Focal adhesion	[45–47]	7.75E-03	**Count 14:** LAMA2, ARHGAP5, ROCK1, MYLK3, COL27A1, ITGB5, ITGA2, PRKCG, LAMC2, HGF, PIK3R3, LAMB1, CRK, SHC4
	hsa04015:Rap1 signaling pathway	[48]	9.08E-03	**Count 14:** GRIN2A, SIPA1L3, PRKCG, HGF, APBB1IP, RALGDS, PRKD1, PLCB4, ADCY9, TEK, RAPGEF4, PIK3R3, CRK, ANGPT4
	hsa04611:Platelet activation	[49]	3.80E-02	**Count 9:** PLCB4, ROCK1, ADCY9, LYN, MYLK3, COL27A1, ITGA2, PIK3R3, APBB1IP
	hsa04020:Calcium signaling pathway	[50,51]	3.87E-02	**Count 11:** GRM5, PLCB4, ADCY9, MYLK3, PHKA1, GRIN2A, RYR1, PPP3CC, PRKCG, HTR2C, HTR5A
	hsa04151:PI3K-Akt signaling pathway	[41,52–54]	4.81E-02	**Count 17:** PHLPP2, CSH2, IFNA10, ITGB5, ITGA2, HGF, CDC37, LAMA2, COL27A1, TEK, GYS1, PIK3AP1, LAMC2, LAMB1, PPP2R2B, PIK3R3, ANGPT4
	hsa04915:Estrogen signaling pathway	[55,56]	7.18E-02	**Count 7:** PLCB4, ADCY9, FKBP4, GPER1, PIK3R3, MMP2, SHC4
Inception-v3	hsa04151:PI3K-Akt signaling pathway	[41,52–54]	4.12E-02	**Count 24:** EGFR, HRAS, PHLPP2, HSP90AA1, COL3A1, ITGB5, HGF, IGF1R, VWF, LAMA4, CDKN1B, COL6A5, CHRM1, COL27A1, ITGA8, IFNA4, ITGA7, GNB5, PIK3AP1, LAMB1, PPP2R2C, FGF3, IFNA17, FGF4
	hsa04512:ECM-receptor interaction	[57,58]	4.36E-02	**Count 9:** VWF, LAMA4, COL6A5, COL27A1, ITGA8, COL3A1, ITGA7, ITGB5, LAMB1
	hsa04740:Olfactory transduction	[44]	6.30E-02	**Count 26:** OR2A25, OR5L1, OR1J1, OR5L2, OR4A5, OR52D1, OR4C3, OR9Q2, OR13C5, OR10G8, OR5B17, OR52R1, OR2S2, OR2A5, OR5R1, OR1K1, OR8G5, OR7A5, OR9G4, OR2AE1, OR5T3, OR5M8, OR4X2, OR4A15, OR13D1, OR51A7
	hsa04114:Oocyte meiosis	[58]	6.53E-02	**Count 10:** PGR, PLCZ1, ANAPC2, IGF1R, SLK, ANAPC5, PLK1, CPEB3, FBXW11, ITPR2
	hsa01212:Fatty acid metabolism	[59]	6.56E-02	**Count 6:** ACADS, EHHADH, FADS2, ACAT1, ACSBG1, ACOX3
	hsa04144:Endocytosis	[47]	8.09E-02	**Count 17:** EGFR, HRAS, KIF5B, KIF5A, KIAA0196, VPS37B, ARFGEF1, IGF1R, SH3GLB2, FOLR1, WWP1, ARPC5L, ZFYVE16, SPG20, AGAP3, ARAP1, HSPA8
	hsa04510:Focal adhesion	[45–47]	8.60E-02	**Count 15:** EGFR, HRAS, COL3A1, ITGB5, HGF, VAV2, BIRC2, IGF1R, VWF, LAMA4, COL6A5, COL27A1, ITGA8, ITGA7, LAMB1
ResNet-50	hsa04512:ECM-receptor interaction	[57,58]	4.86E-02	**Count 9:** GP5, COL6A3, HSPG2, ITGA10, AGRN, LAMC1, COL5A2, THBS3, FN1
	hsa04740:Olfactory transduction	[44]	4.95E-02	**Count 27:** OR2AK2, OR10A6, OR2T2, OR10T2, OR10G4, OR2L2, OR2G2, OR2T10, OR2G6, OR2T6, OR10R2, OR6N1, OR2M7, OR5P3, OR2B11, OR11L1, OR10J3, OR2M2, OR2M3, OR4D6, OR10Z1, OR2T27, OR6K3, OR6K2, OR6K6, OR10K1, OR10K2
	hsa04370:VEGF signaling pathway	[60,61]	6.30E-02	**Count 7:** SH2D2A, PLA2G4A, PIK3CD, PLCG2, BAD, MAPKAPK2, PIK3R3
	hsa04070:Phosphatidylinositol signaling system, including Endocytosis, PI3K-Akt, and Focal adhesion pathways		8.51E-02	**Count 9:** PIK3C2B, PIK3CD, PLCG2, ITPKB, PIP5K1A, PI4KB, CDS1, PIK3R3, INPP5B
Inception-ResNet-v2	hsa05202:Transcriptional misregulation in cancer	[62–65]	2.61E-02	**Count 15:** KMT2A, CEBPE, RXRB, RELA, MET, AFF1, MMP3, ITGAM, ATM, SS18, TAF15, HOXA10, HIST1H3G, PLAU, MLLT3
	hsa04151:PI3K-Akt signaling pathway	[41,52–54]	3.57E-02	**Count 25:** PPP2R1B, HSP90AB1, COL4A2, RELA, MET, ITGB4, ITGA1, FGF10, ITGA2, LPAR1, PCK2, CCNE2, LAMA2, GH1, ITGA9, EIF4E, COL6A6, PRLR, IFNA4, RAC1, TEK, CREB3L2, EFNA5, PPP2R2B, ANGPT2
	hsa04614:Renin-angiotensin system	[66–68]	9.21E-02	**Count 4:** AGTR1, ACE, CMA1, CTSG

Writing - original draft. **Sen Li:** Formal analysis, Methodology. **Yang Zhang:** Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing - original draft.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2021.01.010.

References

[1] DeBerardinis RJ, Lum JJ, Hatzivassiliou G, et al. The biology of cancer: metabolic reprogramming fuels cell growth and proliferation. Cell Metab 2008;7:11–20.
[2] Yau EH, Kummetha IR, Lichinchi G, et al. Genome-wide CRISPR screen for essential cell growth mediators in mutant KRAS colorectal cancers. Cancer Res 2017;77:6330–9.
[3] Bray F, Ferlay J, Soerjomataram I, et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018;68:394–424.
[4] Chen M, Zhang B, Topatana W, et al. Classification and mutation prediction based on histopathology H&E images in liver cancer using deep learning. npj Precis Oncol 2020;4:1–7.
[5] Jiao W, Atwal G, Polak P, et al. A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. Nat Commun 2020;11:1–12.
[6] Keller L, Belloum Y, Wikman H, et al. Clinical relevance of blood-based ctDNA analysis: mutation detection and beyond. Br J Cancer 2020:1–14.
[7] Zehir A, Benayed R, Shah RH, et al. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. Nat Med 2017;23:703.

[8] Abdel-Basset M, El-Shahat D, El-henawy I, et al. A new fusion of grey wolf optimizer algorithm with a two-phase mutation for feature selection. Expert Syst Appl 2020;139:112824.
[9] Sundaram L, Gao H, Padigepati SR, et al. Predicting the clinical impact of human mutation with deep neural networks. Nat Genet 2018;50:1161–70.
[10] Liu X, Li L, Peng L, et al. Predicting cancer tissue-of-origin by a machine learning method using DNA somatic mutation data. Front Genet 2020;11:674.
[11] He B, Lang J, Wang B, et al. TOOme: a novel computational framework to infer cancer tissue-of-origin by integrating both gene mutation and expression. Front Bioeng Biotechnol 2020:8.
[12] Ismael SAA, Mohammed A, Hefny H. An enhanced deep learning approach for brain cancer MRI images classification using residual networks. Artif Intell Med 2020;102:101779.
[13] Chang P, Grinband J, Weinberg B, et al. Deep-learning convolutional neural networks accurately classify genetic mutations in gliomas. Am J Neuroradiol 2018;39:1201–7.
[14] Lee S-I, Celik S, Logsdon BA, et al. A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia. Nat Commun 2018;9:42.
[15] Kumar A, Singh SK, Saxena S, et al. Deep feature learning for histopathological image classification of canine mammary tumors and human breast cancer. Inf Sci 2020;508:405–21.
[16] Chen R, Yang L, Goodison S, et al. Deep-learning approach to identifying cancer subtypes using high-dimensional genomic data. Bioinformatics 2020;36:1476–83.
[17] van IJzendoorn DG, Szuhai K, Briaire-de Bruijn IH, et al. Machine learning analysis of gene expression data reveals novel diagnostic and prognostic biomarkers and identifies therapeutic targets for soft tissue sarcomas. PLoS Comput Biol 2019;15:e1006826.
[18] Coudray N, Ocampo PS, Sakellaropoulos T, et al. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nat Med 2018;24:1559–67.
[19] Sun Y, Zhu S, Ma K, et al. Identification of 12 cancer types through genome deep learning. Sci Rep 2019;9:1–9.
[20] Yuan Y, Shi Y, Li C, et al. DeepGene: an advanced cancer type classifier based on deep learning and somatic point mutations. BMC Bioinf 2016;17:476.
[21] AlShibli A, Mathkour H. A shallow convolutional learning network for classification of cancers based on copy number variations. Sensors 2019;19:4207.
[22] Selvaraju RR, Cogswell M, Das A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. p. 618–26.
[23] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 3rd International Conference on Learning Representations, 2015.
[24] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. p. 2818–26.
[25] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. p. 770–8.
[26] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning. Thirty-first AAAI conference on artificial intelligence, 2017.
[27] Hosmer Jr DW, Lemeshow S, Sturdivant RX. Applied logistic regression. John Wiley & Sons; 2013.
[28] Bayes T. An essay towards solving a problem in the doctrine of chances. 1763, MD computing: computers in medical practice 1991;8:157..
[29] Hart P. The condensed nearest neighbor rule (Corresp.). IEEE Trans Inf Theory 1968;14:515–6.
[30] Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. J Am Stat Assoc 2002;97:77–87.
[31] Cortes C, Vapnik V. Support-vector networks. Machine Learning 1995;20:273–97.
[32] Burges CJ. A tutorial on support vector machines for pattern recognition. Data Min Knowl Disc 1998;2:121–67.
[33] Breiman L. Random forests. Machine Learning 2001;45:5–32.
[34] Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat 2001;29:1189–232.
[35] Hasan MA, Lonardi S. mClass: Cancer Type Classification with Somatic Point Mutation Data. In RECOMB International conference on Comparative Genomics. 2018, p. 131-145. Springer..
[36] Springenberg JT, Dosovitskiy A, Brox T, et al. Striving for simplicity: the all convolutional net. 3rd International Conference on Learning Representations, 2015.
[37] Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature 1986;323:533–6.
[38] 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. p. 89–96.
[39] Lvd Maaten, Hinton G. Visualizing data using t-SNE. J Mach Learning Res 2008;9:2579–605.
[40] Huang DW, Sherman BT, Tan Q, et al. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucl Acids Res 2007;35:W169–75.
[41] Shukla S, MacLennan GT, Hartman DJ, et al. Activation of PI3K-Akt signaling pathway promotes prostate cancer cell invasion. Int J Cancer 2007;121:1424–32.
[42] Taylor BS, Schultz N, Hieronymus H, et al. Integrative genomic profiling of human prostate cancer. Cancer Cell 2010;18:11–22.
[43] Sanz G, Leray I, Dewaele A, et al. Promotion of cancer cell invasiveness and metastasis emergence caused by olfactory receptor stimulation. PLoS One 2014:9.
[44] Weber L, Maßberg D, Becker C, et al. Olfactory receptors as biomarkers in human breast carcinoma tissues. Front Oncol 2018;8:33.
[45] McLean GW, Carragher NO, Avizienyte E, et al. The role of focal-adhesion kinase in cancer—a new therapeutic opportunity. Nat Rev Cancer 2005;5:505–15.
[46] Johnson TR, Khandrika L, Kumar B, et al. Focal adhesion kinase controls aggressive phenotype of androgen-independent prostate cancer. Mol Cancer Res 2008;6:1639–48.
[47] Fan S, Liang Z, Gao Z, et al. Identification of the key genes and pathways in prostate cancer. Oncol Lett 2018;16:6663–9.
[48] Bailey CL, Kelly P, Casey PJ. Activation of Rap1 promotes prostate cancer metastasis. Cancer Res 2009;69:4962–8.
[49] Caine GJ, Lip GY, Stonelake PS, et al. Platelet activation, coagulation and angiogenesis in breast and prostate carcinoma. Thromb Haemost 2004;92:185–90.
[50] Lin J, Denmeade S, Carducci MA. HIF-1α and calcium signaling as targets for treatment of prostate cancer by cardiac glycosides. Curr Cancer Drug Targets 2009;9:881–7.
[51] Wasilenko WJ, Cooper J, Palad AJ, et al. Calcium signaling in prostate cancer cells: evidence for multiple receptors and enhanced sensitivity to bombesin/GRP. Prostate 1997;30:167–73.
[52] Sarker D, Reid AH, Yap TA, et al. Targeting the PI3K/AKT pathway for the treatment of prostate cancer. Clin Cancer Res 2009;15:4799–805.
[53] Morgan TM, Koreckij TD, Corey E. Targeted therapy for advanced prostate cancer: inhibition of the PI3K/Akt/mTOR pathway. Curr Cancer Drug Targets 2009;9:237–49.
[54] Gao N, Zhang Z, Jiang B-H, et al. Role of PI3K/AKT/mTOR signaling in the cell cycle progression of human prostate cancer. Biochem Biophys Res Commun 2003;310:1124–32.
[55] Lafront C, Germain L, Weidmann C, et al. A systematic study of the impact of estrogens and selective estrogen receptor modulators on prostate cancer cell proliferation. Sci Rep 2020;10:1–12.
[56] Bonkhoff H. Estrogen receptor signaling in prostate cancer: implications for carcinogenesis and tumor progression. Prostate 2018;78:2–10.
[57] Stewart DA, Cooper CR, Sikes RA. Changes in extracellular matrix (ECM) and ECM-associated proteins in the metastatic progression of prostate cancer. Reprod Biol Endocrinol 2004;2:2.
[58] Guo L, Lin M, Cheng Z, et al. Identification of key genes and multiple molecular pathways of metastatic process in prostate cancer. PeerJ 2019;7:e7899.
[59] Berquin IM, Edwards IJ, Kridel SJ, et al. Polyunsaturated fatty acid metabolism in prostate cancer. Cancer Metastasis Rev 2011;30:295–309.
[60] Goel HL, Mercurio AM. VEGF targets the tumour cell. Nat Rev Cancer 2013;13:871–82.
[61] Goel HL, Chang C, Pursell B, et al. VEGF/neuropilin-2 regulation of Bmi-1 and consequent repression of IGF-IR define a novel mechanism of aggressive prostate cancer. Cancer Discov 2012;2:906–21.
[62] Yu J, Yu J, Mani R-S, et al. An integrated network of androgen receptor, polycomb, and TMPRSS2-ERG gene fusions in prostate cancer progression. Cancer Cell 2010;17:443–54.
[63] Kumar-Sinha C, Tomlins SA, Chinnaiyan AM. Recurrent gene fusions in prostate cancer. Nat Rev Cancer 2008;8:497–511.
[64] Leshem O, Madar S, Kogan-Sakin I, et al. TMPRSS2/ERG promotes epithelial to mesenchymal transition through the ZEB1/ZEB2 axis in a prostate cancer model. PLoS One 2011;6.
[65] Cai C, Hsieh C-L, Omwancha J, et al. ETV1 is a novel androgen receptor-regulated gene that mediates prostate cancer cell invasion. Mol Endocrinol 2007;21:1835–46.
[66] Uemura H, Hasumi H, Ishiguro H, et al. Renin-angiotensin system is an important factor in hormone refractory prostate cancer. Prostate 2006;66:822–30.
[67] Chow L, Rezmann L, Catt K, et al. Role of the renin–angiotensin system in prostate cancer. Mol Cell Endocrinol 2009;302:219–29.
[68] Uemura H, Hoshino K, Kubota Y. Engagement of renin-angiotensin system in prostate cancer. Curr Cancer Drug Targets 2011;11:442–50.