A Self-attention Graph Convolutional Network for Precision Multi-tumour Early Diagnostics with DNA Methylation Data

X ue Jiang
Shanghai Jiao Tong University https://orcid.org/0000-0003-2109-8660
Zhiqi Li
Shanghai Jiao Tong University https://orcid.org/0000-0001-9108-2045
Heng Wang
Shanghai Institute of Technology https://orcid.org/0000-0001-8269-0743
Yanyi Chu
Shanghai Jiao Tong University
Qiankun Wang
Shanghai Jiao Tong University https://orcid.org/0000-0003-2465-1468
Xueying Mao
Shanghai Jiao Tong University
Jing Zhao
Shanghai Jiao Tong University https://orcid.org/0000-0002-9944-3244
Mingming Jiang
Shanghai Jiao Tong University
Bowen Zhao
Shanghai Jiao Tong University https://orcid.org/0000-0001-7421-0869
Yi Xiong
Shanghai Jiao Tong University
Edwin Wang
University of Calgary
Dongqing Wei (dqwei@sjtu.edu.cn)
Shanghai Jiao Tong University https://orcid.org/0000-0003-4200-7502

Research Article

Keywords: Graph convolutional network, Attention mechanism, Deep learning, Cancer-specific methylation patterns, Precision multi-tumour early diagnostics

Posted Date: February 11th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1348334/v1

A Self-attention Graph Convolutional Network for Precision Multi-tumour Early Diagnostics with DNA Methylation Data

Xue Jiang[#], Zhiqi Li[#], Heng Wang, Yanyi Chu, Qiankun Wang, Xueying Mao, Jing Zhao, Mingming Jiang, Bowen Zhao, Yi Xiong, Edwin Wang, Dongqing Wei*

Abstract

DNA methylation data-based precision tumour early diagnostics is emerging as the state of the art for molecular tumour recognition, which could capture the signals of cancer occurrence 3~5 years in advance and clinically more homogenous groups. However, the sensitive of early detection for many tumors is about 30%, which needs to be greatly improved. Nevertheless, on the basis of the whole genome bisulfite sequencing methylation data, a comprehensive characterisation of the entire molecular genetic landscape of the tumor as well as the subtle differences between different tumours could be identified. With the accumulation of methylation data, high performance deep learning models that considering and modeling more unbiased information need to be developed. According to the above analysis, we designed a pipeline to investigate genome-wide DNA methylation patterns for precision multi-tumour early diagnostics. We proposed a graph convolutional network considering the attention mechanism to dissect DNA methylation heterogeneity of different cancer types. The attention mechanism in the graph convolutional network architecture could detect the cancer signals from genome-wide methylated molecular interactions, improving the robustness and sensitivity of multi-tumor classification. Experimental results demonstrates the feasibility of precision multi-tumour early diagnostics with the DNA methylation data. The workflow presented here is very useful for tumor classification, which highly relevant for the future blood diagnosis and treatment of the tumour.

Keywords: Graph convolutional network, Attention mechanism, Deep learning, Cancer-specific methylation patterns, Precision multi-tumour early diagnostics

1. Introduction

The world health organization (WHO) of the internal agency for research on cancer (IARC) states that there are 18.1 million cancer cases appeared and 9.6 million deaths for cancer in 2018 [1, 2]. The world's cancer situation is grim. With the aging and rapid growth of the population, the incidence of cancer and the number of deaths continue to rise. Cancer will become the leading cause of death in the 21st century and will be a major barrier to improving life expectancy in countries around the world. While much of current cancer researches are focused on developing new therapeutics, studies have shown that early detection has the potential to reduce both treatment cost and mortality rates from cancer by a significantly amount.

Early cancer diagnosis offers the opportunity to identify tumours when cures are more achievable, outcomes are superior, and treatment can be less morbid. However, most cancer types currently lack an effective non-invasive early screening option [3]. Effective screening paradigms exist only for a small subset of cancers. Usually, cancer diagnoses are often prompted by symptoms. But, many cancers do not cause the appearance of symptoms until late in disease development. Recently, DNA methylation data-based cancer diagnostics are currently emerging as the state of the art for molecular tumour identification [4-7].

DNA methylation is a primary epigenetic mechanism that involves the covalent modification of cytosine bases by the addition of a methyl group at the 5' position. DNA methylation plays a critical role in the normal development and regulation of many cellular processes [8], with DNA methylation profiles being cell type and tissue specific [9]. DNA methylation occures before tumour and is an important mechanism of it. A global loss of DNA methylation (hypomethylation) accompanied by increased de novo DNA methylation (hypermethylation) of CpG-rich regions can be implicated in many tumors [10, 11]. In recent years, DNA methylation (5-methylcytosine, 5mC) in cell-free DNA (cfDNA) have been successfully detected in clinical samples by a range of genome-wide approaches, demonstrating high clinical potential in cancer early diagnosis, prognosis and/or treatment response. Resource competition between tumour cells and normal cells could result in cell death. The DNA of died cells were further released by them and were circulated in the blood. Therefore methylation changes of circulating tumour DNA (ctDNA) can be detected in the blood to enable early diagnosis of tumour 3~5 years (ultra early) before it formation.

Precision early tumour diagnosis is crucial for cancer patients. Studies have shown that early-detection of cancer can significantly increase a patient's 10-year survival rate (from 30% to 90%). The 5-year survival rate for patients with early-stage cancer can be up to 91% if treated with early intervention[12-15]. DNA methylation is particularly suitable for individualized early cancer diagnostics for three reasons: 1) large amount of researchers have identified characteristics differences in the DNA methylation profiles between subgroups of patients using DNA methylation data [16-19], demonstrating the power of DNA methylation data for analyzing epigenetic heterogeneity, 2) since the cancer methylome is a combination of both somatically acquired DNA methylation changes and characteristics reflecting the cell of origin, it is especially suitable for molecular classification of tumours and thus for stratifying cancer patients [20-24], 3) it has been convincingly shown that DNA methylation profiling is highly robust and reproducible even from small samples and poor quality material [25]. Although 5mC biomarkers are sensitive for early tumour detection, in large-scale population screening, such multi-cancer detection approach would require high specificity, clinical useful sensitivity, to limit the scope, cost, and complexity of evaluating asymptomatic patients.

Nowadays, the sensitive of the early (stage I) diagnosis of many tumours, such as breast,

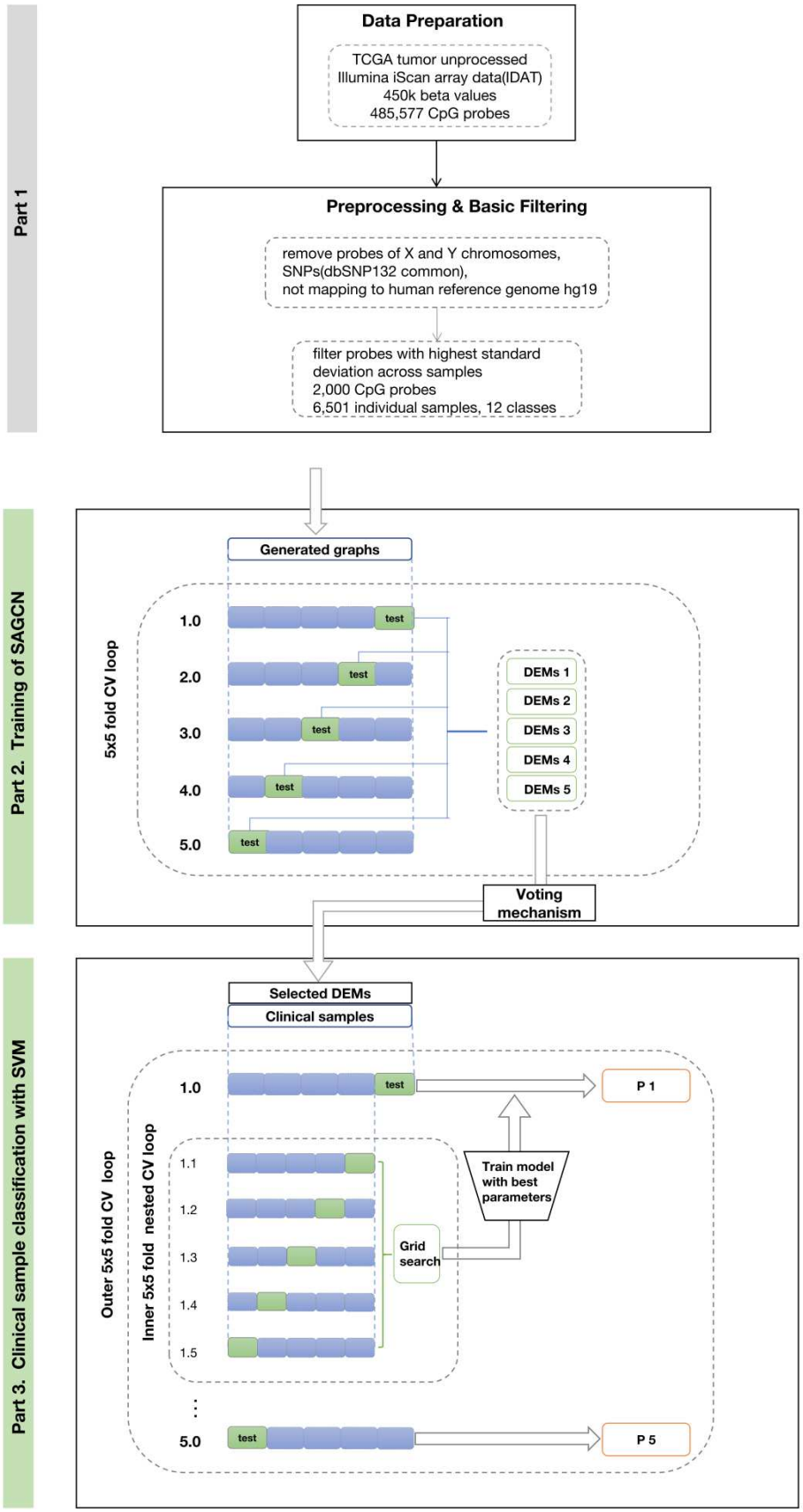
esophageal, kidney, lung, prostate, and uterine cancers, is about 30% with methylation, which needs to be improved greatly. Existing studies of tumour diagnosis based on methylation data mainly include two steps. First, information regions, including differentially methylated CpG sites and cancer-related genomic regions reported in literatures, are determined. Then targeted panel is made for in-depth sequencing. Next, regression models often used to fit the methylation beta value data of the selected sites. As cancer is a disease of the genome, elements on the genome have complex regulatory relationships. This kind of dimension reduction and then conduct classification using linear models would ignore large amount of the dynamic interactions between molecules and restrict the understanding towards disease pathological mechanism. Moreover, the classification accuracy of the model depends heavily on the selected marker methylated sites, which is based on the statistical significant difference in beta values, lacking the interpretability of the data and limiting the applicability to non-labeled data sets. As a result of the development of high-throughput whole genome bisulfite sequencing (WGBS) technology, it is no longer difficult to obtain genome-wide methylation information. However, the existing bioinformatic tools have not fully utilized this information to distinguish different cancers and explain the pathological mechanisms. For many cancers, the early diagnosis sensitivity needs to be greatly improved. The advanced graphical neural network models in the field of artificial intelligence present opportunities and challenges for the analysis of the whole genome scale molecular data, improving the robustness and accuracy of the model as well as increasing understanding towards the pathological mechanism. Therefore, a methodology for DNA methylation analysis, which could use whole genome information effectively, is urgently needed.

Based on the above analysis, a graph convolutional network model was designed to distinguish eleven most common types of cancer with WGBS methylation data. The graph convolutional network, on the one hand, could model the information of the whole genome methylation sites, improving the stability and accuracy of prediction while the sequencing depth is insufficient; on the other hand, could increase the explanatory ability of differentiating the pathological mechanism of different cancers. The graph convolutional network was used to model the relationship between molecular regulation. Finally, the attention mechanism was used to learn signals formed by the molecular interaction, capture the weak pathological signals in the early stage, thus to prioritize disease CpG sites. Collectively, the framework proposed here advanced multi-tumour diagnostics, with robust and higher accuracy of early tumour diagnosis.

2. Materials and Methods

2.1 Study design and data preparation

Based on the graph convolutional network we present a pipeline to conduct feature selection and sample classification with the DNA methylation data. The pipeline included four parts. Part 1 is data preparation, preprocessing and basic filtering. Part 2 is the training process with self-attention graph convolutional network (SAGCN). Part 3 is clinical sample classification with support vector machine (SVM). Part 4 is performance evaluation. The details of the framework is show in Figure 1.


Part 1
Data Preparation
TCGA tumor unprocessed
Illumina iScan array data(IDAT)
450k beta values
485,577 CpG probes
Preprocessing & Basic Filtering
remove probes of X and Y chromosomes,
SNPs(dbSNP132 common),
not mapping to human reference genome hg19
filter probes with highest standard
deviation across samples
2,000 CpG probes
6,501 individual samples, 12 classes
Part 2. Training of SAGCN
Generated graphs
5x5 fold CV loop
1.0
2.0
3.0
4.0
5.0
test
DEMs 1
DEMs 2
DEMs 3
DEMs 4
DEMs 5
Voting mechanism
Part 3. Clinical sample classification with SVM
Selected DEMs
Clinical samples
Outer 5x5 fold CV loop
Inner 5x5 fold nested CV loop
1.1
1.2
1.3
1.4
1.5
Grid search
Train model with best parameters
P 1
P 5

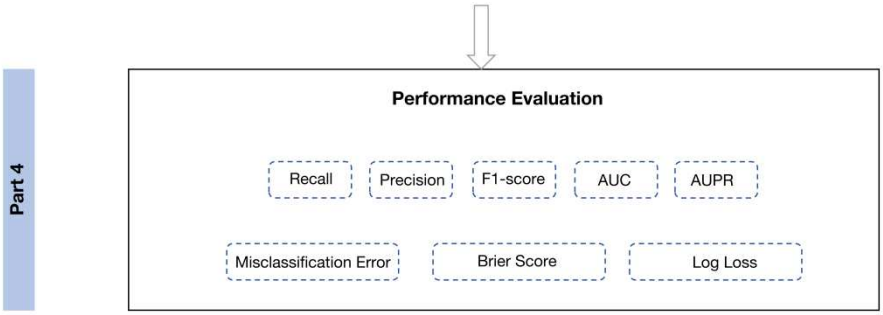


Fig. 1. **The pipeline of the self-attention graph convolutional network based computational framework.** Data preparation and pre-processing steps are described in Part 1. The training process of self-attention graph convolutional network with a 5-fold cross validation is shown in Part 2. The number '1,2,...,5' within these boxes indicate the fold identification numbers of the fivefold CV. Light blue and rectangles indicate the training and test sets, respectively. Clinical sample classification with support vector machine (SVM) is shown in Part 3. Light-blue and green boxes in the outer CV loop represent the outerfold (1.0; 2.0; ...5.0) training and test sets, respectively. Similary, in the inner 5x5 fold nested CV loop (1.1, 1.2, ..., 1.5), light blue and rectangles indicate the nested training and test sets, respectively. The corresponding calibration set was test set of outer loop in order to prevent information leakage. The collection of green rectangles (s1.1-1.5) represent the combined calibration set of raw probability outputs ('raw scores') that is used for training post-processing, that is the fitting of the tuned calibrator algorithm on the outerfold 'raw scores', to generate calibrated probabilities (orange boxes, P1.0-5.0).

From The Cancer Genome Atlas (TCGA) NCI GDC (National Cancer Institute Genomic Data Commons) Repository (https://portal.gdc.cancer.gov/repository), we collected 450k DNA methylation data of 6557 samples, including 11 types of tumour primary site, such as bladder, brain, breast, bronchus and lung, cervix uteri, corpus uteri, kidney, liver and intrahepatic bile ducts, prostate gland, stomach, and thyroid gland. The samples included 5905 primary tumour sites, 596 solid tissue normal samples, 37 recurrent tumour samples, 17 metastatic samples, 2 new primary samples. Recurrent tumour samples, metastatic samples and new primary samples were excluded. Primary tumour samples and solid tissue normal samples were used for further analysis. In this study, the samples were divided into 12 classes, all solid tissue normal samples and 11 types of tumour samples. The distribution of these samples is shown in supplementary Figure S1.

The detailed patient demographics and clinical characteristics can be found in supplementary Table 1. For details on how to prepare the 450k DNA methylation tumour samples from TCGA, please see the GitHub repository (https://github.com/lizhiqi0506/GNN_CancerPreDiagnosisWithMeth), and to download the source data, visit the NCI GDC Legacy Archive (https://gdc-portal.nci.njh.gov/legacy-archive).

DNA methylation data is based on genome-wide quantitative measurements of DNA methylation at 485,577 CpG sites using Illumina Human Methylation 450 BeadChip technologies (450k; Illumina). Illumina Human Methylation BeadChip arrays are a popular tool to measure genome-wide single-nucleotide CpG site methylation levels [26]. The 450k BeadChip provides >98% coverage of reference sequence genes and 96% of CpG islands. The beta value represents the

average methylation fraction (AMF), which is computed for each genomic region by summing the number of observed cytosines at all covered CpG sites and dividing by the total sequencing depth at all covered CpG sites in each region. For each methylation locus, the amount of methylated DNA is denoted as Meth, and the amount of unmethylated DNA is denoted as Unmeth. The beta value is computed as Meth/(Unmeth + Meth), and further used in downstream analysis.

The matrix $M = [m_{ij}] \in R^{N_n \times N_s}$ is used to represent the DNA methylation data, in which columns represent samples and rows represent methylation sites. m_{ij} represents the beta value of the i th methylation site in the j th sample. To reduce the complexity of downstream analysis and improve the performance of the computational method, pre-processing and basic filtering steps are needed.

Step 1. filter out samples with uncompleted site covers.

Step2.remove probes of X and Y chromosomes, probes containing single-nucleotide polymorphisms (dbSNP132Common), and probes not mapping uniquely to human reference genome 19. Finally 396,013 methylation sites were left.

Step 3. 2000 most variably methylated probes were further selected according to the highest standard deviation across all samples.

Step 4. then, a principal component analysis (PCA) were conducted. The first 100 principal components were then used as input data for the t-Distributed Stochastic Neighbour Embedding (t-SNE, Rtsne package version 0.11, https://github.com/mwsill/mnp_training/blob/master/tsne.R) for clustering analysis [27, 28]. The cluster results is shown in Figure 2.

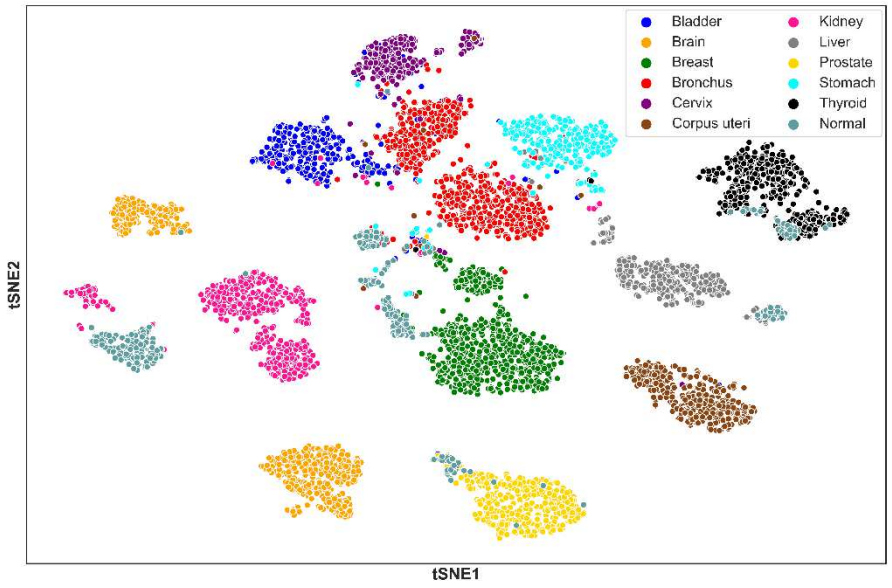


Fig. 2. t-**SNE visualization of 6501 samples analyzed by PCA, colored by clusters, and labeled by the sample type.**

2.2 Self-attention graph convolutional network

2.2.1 Construction of methylation sites interaction graph

In this study, we used weighted gene co-expression network analysis (WGCNA) to construct methylation sites interaction graph. Let $G(V, A, X)$ represent a methylation sites interaction graph of one class, where V represents the methylation sites in the network, $A = [a_{ij}] \in R^{N_m \times N_m}$ is the adjacency matrix, $X = [x_{ij}] \in R^{N_m \times N_s}$ is the original feature matrix, N_m is the number of methylation sites, N_s is the number of samples. It should be noted that, in the network construction, the interaction relationships between methylation sites was determined by the following criterion:

$$C(i, j) = [corrcoef(X_i., X_j.)]^{\rho}$$

where $corrcoef(.,.)$ is a function that calculates the Pearson correlation coefficient between edges. We set a threshold 0.01 to determine whether there is interaction between two methylation sites.

$$a_{ij} = \begin{cases} 1, & C(i, j) > 0.01, \\ 0, & C(i, j) \leq 0.01. \end{cases}$$

To deal with small sample size and label imbalance problem, we used a subsampling strategy before the model training. For each class, we randomly sampled 200 samples from corresponding dataset with replacement. Through the above strategy, we created 80 graphs for each class (totally 80 x 12 = 960 graphs). We tackle the task as a graph classification problem.

2.2.2 Graph convolutional network

Deep learning has made great progress in the field of computer vision and natural language processing. Meanwhile, in the real world, most issues can't be described by Euclid field data, but on the contrary can be abstracted as graphs easily. In order to performing deep learning based on graph data, different kinds of graph neural networks (GNNs) were presented. Graph convolutional network (GCN) is one of the most widely used GNNs. Graph convolutional operation can be defined in either the spectral or non-spectral domain. Spectral approaches perform convolution on graph by redefining the convolution operation in the Fourier domain. The classical spectral graph convolution method is formulated as bellow:

$$Z = \sigma(D^{-\frac{1}{2}} A D^{-\frac{1}{2}} X W)$$

Where $X = [x_{ij}] \in R^{N_m \times N_s}$ is the input feature matrix,N_m is the number of methylation sites and N_s is the number of samples, $W = [w_{ij}] \in R^{N_s \times N_f}$ represents weight matrix, N_f is the updated dimensionality of output features, σ is a nonlinear activation function, $Z \in R^{N_m \times N_f}$ is feature matrix after convolution, A represents adjacency matrix for methylation sites interaction graph, D is the diagonal matrix of A with $D_{ii} = \sum_j A_{ij}$.

2.2.3 Self-attention graph pooling

Attention mechanisms helps to focus on those most important features and lose sight on those unimportant features, which is very useful in the task of classification and data mining. In this study, we proposed a graph pooling method with attention mechanism (SAGPooling). The attention mechanism is realized by graph convolution. For a methylation sites interaction graph with N_m nodes, the self-attention score $Z_{att} \in R^{N_m \times 1}$ is defined and calculated as

$$Z_{att} = \sigma(D^{-\frac{1}{2}} A D^{-\frac{1}{2}} X W_{att})$$

Where σ is the activation function (e.g. ReLU)，$W_{att} \in R^{N_f \times 1}$ is the only parameter of SAGPooling layer. By utilizing graph convolution to obtain self-attention scores, both graph patterns and topology are considered. During the step, the hyperparameter of pooling ratio $k \in (0, 1)$ is used to determine the nodes selected. We ranked the nodes in descending order based on the value of Z_{att} and top kN_m nodes are selected.

$$idx = TopRank(Z_{att}[kN])$$

Where $TopRank$ is function that returns the indices of the top $[kN]$ values. For a input graph $G(V, A, X)$, the output graph $G'(V', A', X')$ can be obtained as:

$$V' = V[idx]$$

$$A' = A[idx][idx]$$

$$X' = X[idx] \odot Z_{att}[idx]$$

Where $[idx]$ is an indexing operation, $\odot$ is the broadcasted elementwise product.

2.2.4 Self-attention graph convolutional network model

We proposed a graph neural network mixed with GCN layers and SAGPooling layers (SAGCN). The structure of SAGCN can be divided into two part. The first part used for node selection and graph embedding. It contains 3 sequential modules, with a GCN layer followed by a SAGPooling layer in each module.

The activation function of GCN layer is ReLU. For a input graph $G^{(0)}(V^{(0)}, A^{(0)}, X^{(0)})$, in the first module, we get $G^{(1)}(V^{(0)}, A^{(0)}, X^{(1)})$ after the GCN layer. And then, SAGPooling is performed on $G^{(1)}$:

$$G_p^{(1)}(V^{(1)}, A^{(1)}, X_p^{(1)}) = SAGPooling[G^{(1)}(V^{(0)}, A^{(0)}, X^{(1)}), k_1]$$

Where $SAGPooling[.,.]$ operation of SAGPooling layer and k_1 is the pooling rate of the first module's SAGPooling layer. And then $G_p^{(1)}$ will be the input of the second module. Finally, after all the three modules processing, we get $G_p^{(3)}(V^{(3)}, A^{(3)}, X_p^{(3)})$ with $N_d = N_m k_1 k_2 k_3$ methylation sites left, k_2 and k_3 are the pooling rate of SAGPooling layers of the second and

third module respectively. Finally, $V^{(3)}$ is the DEMs selected from the input graph by the algorithm.

Next, we perform scatter-max and scatter-mean on $G_p^{(3)}$ to get an graph embedding (GE). Scatter-max operation is to pick the maximum element in each row as the feature of corresponding node. Scatter-mean operation is to calculate the average value of each row as the feature of corresponding node. GE is generated by concatenating the two vectors, and then served as the input of the second part.

The second part of the model mainly functions as a classifier, which consists of a CNN layer and a MLP (multilayer perceptron) with one hidden layer. Finally, the Cross Entropy loss function is computed on the output.

$$\sigma(\widehat{y_j}) = \frac{e^{\widehat{y_j}}}{\sum_{k=1}^{K} e^{\widehat{y_k}}}$$

$$Loss = -\sum_{k=1}^{K} y_k ln\sigma(\widehat{y_k})$$

Where $\widehat{y}$ is the output logits of the MLP and K is the number of classes. The SAGCN model is illustrated in Figure 3.

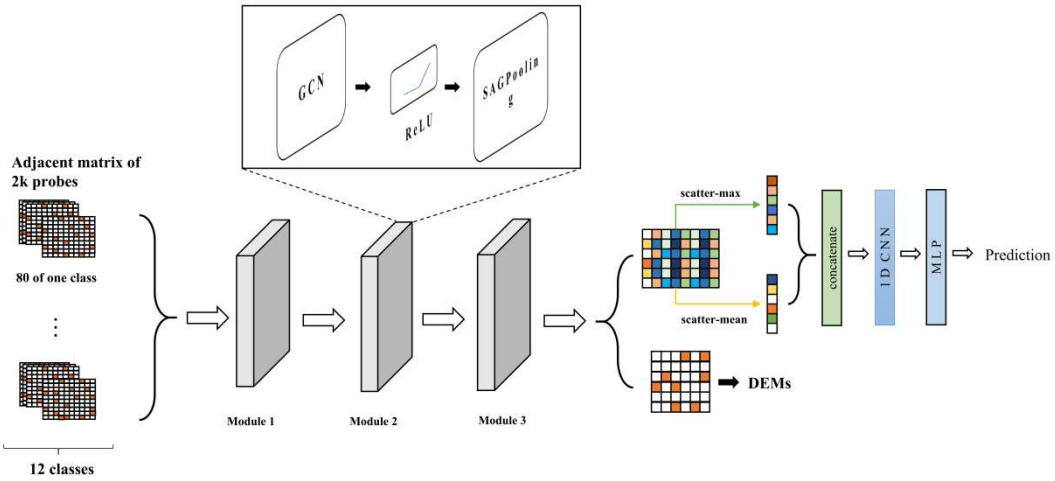


Fig. 3. **The workflow of the proposed SAGCN model.**

2.3 Selecting DEMs

To obtain tissue-specific methylation pattern, we first construct interaction relationships between these methylation sites for each class. Then, those graphs are used to train the SAGCN, in which graph convolutional layer was used to aggregate features and self-attention mechanism (SAGPooling) was used to extract the DEMs.

For each input graph, we get the methylation sites set $V^{(3)}$ from the first part of SAGCN

model. For different input graphs, theoretically there is a minor changes of the selected $V^{(3)}$. Therefore, to get a convergent and robust DEMs set, we utilized a voting mechanism to determine the final selection that show up most frequently for all input graphs.

Algorithm 1. DEMs selection and sample classification with SAGCN

Input: filtered methylation sites L;
Methylation matrix M;
Sample sets $S = \{S_k | k = 1, \ldots, K\}$, S_k represents the samples belong to class k;
The number of samples N_s to create a graph;
The number of graphs N_g for each class;
The edge threshold e.
Output: SAGCN model with trained weights;
DEMs set D;
Calibrated sample class probability;

function Sampler(M, k, N_s, S)
  $S'_k \leftarrow$ N_s samples randomly sampled from S_k with replacement
  $M'_k \leftarrow$ Column induced submatrix of M from S'_k
end function
function Create graph (L, M'_k, e)
  $V \leftarrow L$
  $C \leftarrow$ Correlation coefficent matrix created by WGCNA
  $A \leftarrow$ Adjacency matrix,$A(i,j) = 1$ if $C(i,j) > e$ else 0
  $G(V, A, M'_k) \leftarrow$ the graph created from M'_k
end function
function Create_Dataset (L, M, S, N_g, e)
  $Dataset \leftarrow$ [], $y \leftarrow$ []
  for k in 0:11 **do**
    for i in 1:N_g **do**
      $M'_k \leftarrow$ Sampler (M, k, N_s, S)
      $G \leftarrow$ Create graph (L, M'_k, e)
      $Dataset$.append(G), y.append(k)
    end for
  end for
end function
Run Create Dataset to obtain $N_g \times 12$ graphs and their labels y
$V_D \leftarrow$ []
for each minibatch **do**
  GE, $V^{(3)} \leftarrow$ the result of the first part of SAGCN model
  $\hat{y} \leftarrow$ Forward propagation of GEs
  V_D.append($V^{(3)}$)
  $Loss_{batch} \leftarrow NLLLoss(\hat{y}, y)$
  Update weights through backward propagation
end for
$N_D \leftarrow$ the length of $V^{(3)}$

$D \leftarrow N_D$ Sites that show most frequently in V_D

2.4 Performance evaluation

In this study, eight criterions are used to evaluate the performance of the model, including recall, precision, F1-score, area under receiver operating characteristic curve(AUROC), area under precision-recall curve (AUPR), misclassification error (ME), brier score (BS), and log loss (LL).

The formulas to compute the recall, precision, F1-score, and ME are as below:

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

$$ME = 1 - \frac{TP + TN}{TP + FN + FP + TN}$$

where TP represents true positive rate, TN represents true negative rate, FP represents false positive rate and FN represents false negative rate in the predicting results. Then, the macro averaged recall, macro averaged precision, macro averaged F1-score and macro averaged ME to evaluate the overall model performance of multi-classification. Moreover, the AUC and AUPR for each class and their macro averaged values are also calculated.It should be noted that, the averaged macro evaluation criterion is the average of the evaluation criterion measured in each of the 5x5-fold CV (outer) test sets.

BS is a proper scoring rule that measures the accuracy of probabilistic predictions of mutually exclusive classes. It is applicable to multiclass prediction and is defined as the quadratic difference between the assigned probability and the value (1, 0) for the class:

$$BS = \frac{1}{N_s} \sum_{i=1}^{N_s} \sum_{k=1}^{K} (P_{i,k} - O_{i,k})^2$$

where N_s is the number of samples and K is the number of classes. $P_{i,k}$ is calibrated estimation probability that observation i belongs to class k and $O_{i,k}$ is the actual class of sample i.

$$O_{i,k} = \begin{cases} 1, & if\ sample\ i\ belongs\ to\ class\ k, \\ 0, & otherwise. \end{cases}$$

Besides, LL is extensively used to assess probability estimates of predictor models. BS encourages predicted probabilities and true labels to lie close to each other, whereas LL does not. Extensive empirical testing, however, stressed LL's favorable local property that it will always assign a higher score to a higher probability estimate for correct class. In contrast, BS can perform poorly in this regard. Thus we also used LL to evaluate model performance. A multiclass extension of log loss is shown below:

$$logloss = -logPr(P|O) = -\frac{1}{N_s} \sum_{i=1}^{N_s} \sum_{k=1}^{K} O_{i,k} log(P_{i,k})$$

3. Results

3.1 Performance of SAGCN

Large amounts of experiments have been performed to investigate the performance of the computational framework. Through the pooling rate in the part one of SAGCN, different differentially expressed methylation sites could be extracted, thus to get robust and good classification performance. In the training process, we set learning rate as 0.001, the number of epochs as 60 and the batch size as 5. The early-stopping mechanism is utilized, which means terminating training if there is no performance improvement in consecutive 15 epochs. To evaluate classification performance, 5x5-fold cross validation was implemented. And macro average values of recall, precision, F1-score , AUC, AUPR, ME, BS, and LL are calculated. We hold the pooling rate of the first and the second SAGPooling layer, i.e., k_1 = k_2 = 0.5, and changed the value of k_3 to determine the number of finally selected DEMs, which is denoted as D. By integrating DEMs obtained by each fold in the training process, we used a voting mechanism to selected the most frequently appeared 350 DEMs. The classification performance with different parameter k_3 is shown in Table 2. The sample classification performance of SAGCN + SVM with different k_3 value is shown in Table 3. It is clearly shown that the performance of SAGCN becomes better then worse with the increase of k_3, and it achieves best performance when $k_3 = 0.5$. The recall, precision, F1-score, AUC and AUPR are 0.9917, 0.9918, 0.9917, 0.9998, and 0.9995 respectively. It is noticed that the accuracy is large than 91%, which means that can be applied to clinical practice. The AUC and AUPR are greater than 0.99, indicating that the proposed method has a outstanding ability of classification. The ROC and PR curves with $k_3 = 0.5$ are shown in supplementary Figure S2, demonstrating the outstanding performance of the SAGCN.

Table 2. The performance of SAGCN with different k_3 values

Table 3. The sample classification performance of SAGCN + SVM with different k_3 value

During the training process, we randomly partitioned the samples in the ration of 75/25 as the training/test data, respectively, 10 times. This indicates that we will have 10 different sets of training/test data and each set can yield different selected markers. Each set of training/test data and its result is called an experimental run. In each of the 10 runs, there is a minor changes of the selected $V^{(3)}$ for each cross validation fold, and the majority of these markers were shared by all runs. Therefore, to get a convergent and robust DEMs set, we utilized a voting mechanism to determine the final selection that show up most frequently. We conducted experiments for sample classification with different k_3 values to determine the selected DEMs. The experimental results are shown in Table 3, according to which, we selected 350 DEMs for downstream analysis. The extracted DEMs and related annotations can be seen in supplementary Table 2. To illustrate the effectiveness of selected DEMs, we draw the cluster heatmap (seaborn, python package, version 0.10.1) with the beta value of them, the result is shown in Figure 4, where the rows represent the DEMs and the columns represent samples which are sorted according to class order. It obviously illustrated that different blocks are formed among different classes, indicating the effectiveness of selected DEMs.

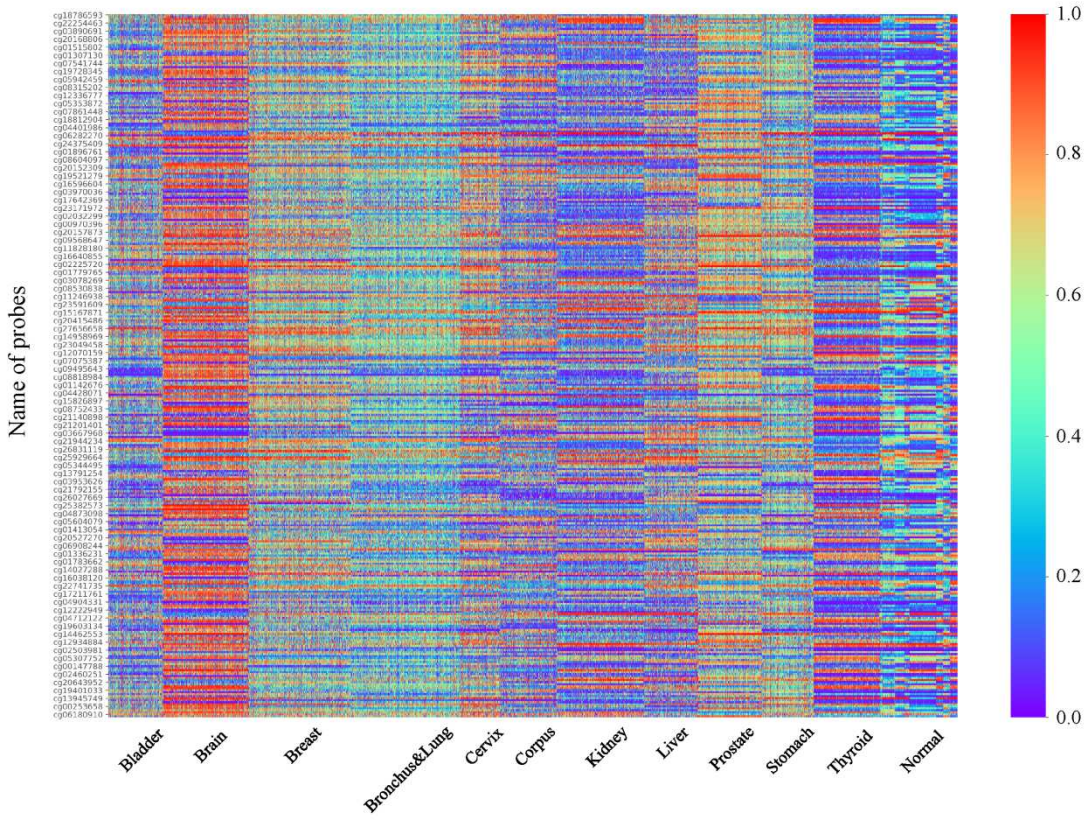


Fig. 4. **The cluster heatmap of the selected DEMs.**

Further more, to test the classification performance of SAGCN for each class with the selected DEMs, the mean values of the basic metrics for each class on 5 folds are shown in Table 4, suggesting the selected DEMs are workable and effectiveness. For the reason of multi-classification, the classification of each class is a binary classification on an unbalanced dataset. In this term, the high accuracy of each class is less believable. Inversely, Recall, Precision, F1-score, AUC, and AUPR can truly reflect the performance of classification. For the 11 kind tumour samples, the evaluation metrics are all greater than 0.95 excepted the Precision of cervix with a value of 0.93, which suggests that nearly 6.73% of the cervix classified as tumor samples are misclassified. Additionally, the Recall, Precision, F1-score, AUC, and AUPR of normal samples are relatively small. Among them, the Precision is only 0.8925, which means that nearly 10.75% normal samples are misclassified as tumour samples. The reason is that, firstly, the normal samples are from different tissues so that they are characterized by tissue-specific DNA methylation features, which may have a intersection with tissue-cancer-specific DNA methylation sites, leading to a high misclassification. Secondly, from Figure 4, we can see that the expression of selected DEMs is closer between Thyroid gland tumour samples and normal samples, which easily cause a misclassification between thyroid gland tumour samples and normal samples.

Table 4. The classification performance of SAGCN for each kind tumors and normal samples, respectively

More importantly, we checked and verified the classification performance of SAGCN + SVM for tumors of stage I and stage II with the selected DEMs. The results are shown in Table 5 and

Table 6, respectively. There are 7 kinds of tumor with stage I and stage II in the dataset, totally 2615 samples. The recalls, i.e., sensitivity, of them are all larger than 0.95, illustrating the effectiveness of the computational pipeline and the selected DEMs. In brief, the SAGCN computational framework can effectively prioritize DEMs, which could get super classification performance on nearly all kinds of tumour samples.

Table 5. The classification performance of SAGCN +SVM for tumors of stage I

Table 6. The classification performance of SAGCN +SVM for tumors of stage II

3.2 Comparisons of SAGCN with classic machine learning models

To illustrate the superiority of the model proposed in this study, we compared it with some classic machine learning algorithms directly based on the original 2000 methylation sites, including random forest classifier (RF), extremely randomized trees (ERT), decision tree (DT), and Gaussian naive Bayes (GNB). The results are shown in Table 7, from which we can see that the performance of SAGCN + SVM is best, indicating the effectiveness of SAGCN and the selected DEMs.

Table 7. The performance comparison of SAGCN + SVM with classic machine learning methods

3.3 Comparisons with the state-of-the-art methods

To further prove the superiority of SAGCN, we compared it with two state-of-the-art methods proposed after 2020. One is convolutional neural network and graph convolutional network-based method for predicting classification (CGATCPred)[29], another is hybrid graph convolutional network for predicting cancer drug response (DeepCDR)[30]. Note that all the following experiments are carried out under the same experimental conditions, including 5-fold cross validation, random seed and data partitioning strategy. The results are shown in Table 8. The three methods can all achieve good performance. The statistical test between SAGCN + SVM and the other two methods is 0.08 and 0.27 for AUC, respectively, indicating the sample classification performance of the proposed model is compare able to the state-of-the-art methods.

Table 8. The performance comparison of SAGCN + SVM with the state-of-the-art methods

3.5 External validation with tissue DNA methylation data

To test the stability and extendibility, we applied this computational framework to external data from NCBI Gene Expression Omnibus (GEO) database, under accession GSE90496 [31], GSE155207 [32], GSE158075 [33], and GSE164988. The brain tumours DNA methylation dataset (GSE90496) includes 2801 samples, meanwhile 294 of these samples (294/2801) were selected out, the disease stage of which is initially operation, the age distributed in range of 25 to 70, and the WHO grade is I, II or III. A fumarate hydratase-deficient renal cell carcinoma dataset (GSE155207) was used as kidney tumour validation. The EPIC array methylation data of 20 primary tumor tissues were selected out to conduct validation experiments. Besides, the Human Methylation450-based DNA methylation analysis of paired samples of bronchoscopic biopsy specimens either from the tumor side or from the contralateral tumor-free bronchus in 37 patients with definite lung cancer diagnosis (GSE158075) were used as validation dataset for bronchus and lung cancer. Moreover, 10 samples with histopathology be diagnosed as adenocarcinoma from

GSE164988 were used as stomach cancer validation.

Performance evaluation of SAGCN on these external validation dataset is shown in Table 9. The classification performance for brain and kidney tumor can achieve 100% precision. Moreover, the classification accuracy for bronchus & lung and stomach tumor can achieve more than 94%. It can also achieve 80% for normal samples. The results indicate that the proposed computational framework are robust, generalize well and can be easily applied to other external methylation datasets to get well-performing classification.

Table 9. The performance evaluation of SAGCN+SVM with external validation dataset of tissue DNA methylation

3.6 External validation with cell-free DNA methylation data

Recent analyses of circulating cell-free DNA suggest that approaches using tumour-specific alterations may provide new opportunities for early diagnosis. Nevertheless, cfDNA whole-genome sequencing can detect abnormal chromosomal changes in patients with cancer but detecting such alterations may be challenging, because the total number of alterations may be scare in individuals with low tumor burden, hence not all patients have detectable changes [34]. In this study, we have developed an end-to-end deep learning framework to realize ultra-sensitive in common cancers. The genome-wide methylation approach has leveraged the discovery of relevant epigenomic signals in human cancer, and these regulatory interactions revealed inherited risk loci for cancer detection. The inherited risk loci often involved in tissue-specific DNA methylation, however, tumors originating from the same organ often aggregated by cancer-type-specific hypermethylation. As genome-wide profiles may reveal differences associated with specific tissues, these patterns may also indicate the tissue source of cfDNA.

Methylation features of cfDNA, such as CpG sites, contain information about recent events in body, which can be used to detect and localize cancer [35-37]. cfDNA molecules from loci carrying tissue-specific methylation patterns can be used to identify cell death in a specific tissue. Consequently, methods based on the identification of CpG sites markers or on the exploration of the methylome could be used to accurately recognize cfDNA released by cells and tissues in various human cancer. Epigenetic approaches, based on the examination of cell-free DNA (cfDNA) methylation patterns or the analysis of cfDNA fragmentation, can detect tumor-derived molecules on a genome-wide scale. However, cfDNA is not uniformly distributed across the genome, but rather is differentially sheared and located depending on chromatin organization, gene expression, tissue of origin, epigenetic marks, and mechanism of release, among other factors [38-41]. Identifying the cell and tissue of origin of the cfDNA fragment could pave the way to localization of cancer of unknown primary with liquid biopsy. Ultimately, decoding the epigenetic and environmental fingerprints of cfDNA could improve the detection of early-stage and other pathologies.

Therefore, we further used a cell-free DNA methylation data from the GEO database, under accession number GSE157272 to conduct external validation. First, blood is collected from the patient individuals. cfDNA is extracted from plasma, processed into sequencing libraries, examined by WGS, mapped to the genome, and analyzed to determine cfDNA fragmentation profiles across the genome. 22 samples from GSE157272 with disease stage as high grade prostatic intra-epithelial neoplasia tissue, indolent prostate cancer tissue, and aggressive prostate cancer tissue were used

to conduct prostate cancer validation. The proposed computational framework SAGCN is used to categorize whether individuals is tumor patient and identify the tumor tissue of origin. Performance evaluation of SAGCN in a 5-fold nested cross validation on the external validation data is shown in Table 10. The results indicate that the proposed computational framework are robust, generalize well and can be easily applied to conduct blood-based non-invasive liquid biopsy.

Table 10. The evaluation performance of SAGCN +SVM with cell free DNA methylation data

4. Conclusion and Discussion

Cancer diagnosis is complex because the majority of malignant tumors present with long periods of latency and lack of clinical presentation at early stages. Nevertheless, epigenetics provides an important molecular link between genetic programming and environmental signals, and such changes in the DNA methylation and the somatic genomic DNA from the tumour tissue of origin are highly consistent in many disease models. Thus, DNA methylation plays an important role in the process of DNA expression and precision tumour early diagnostics, emerging as the state of the art for molecular tumour recognition.

Here, we present a genome-scale computational framework for DNA methylation analysis. First, methylation interaction graphs are properly constructed. Then, a self-attention graph convolutional neural network model is proposed to extract the DEMs based on those graphs, and conduct sample classification based on the extracted methylation sites. To improve the sample classification accuracy and robustness, based on the selected DEMs, we further designed SVM with inner 5-fold nested cross validation for input sample classification. Large amounts experiments have been conducted and have shown that the effectiveness and robustness performance of the model compared with either classic machine learning methods or state of the art methods.

Moreover, traditional approaches have largely been applied in patients with late stage cancers or have used tumor tissue sequencing to guide mutational analyses in the blood. In this study, we provide a broadly applicable approach for non-invasive detection of early stage tumors that may be useful for screening and management of patients with cancer. Analytic validation of tissue and cell-free DNA methylation for early cancer detection and diagnosis, and the clinical utility of candidate epigenetic alterations can be applied to cancer management.

References:

[1]. World Health Organization. Guide to Early Cancer Diagnosis. (2017)

[2]. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2018. CA Cancer J. Clin. 68, 7–30 (2018).

[3]. Shen, S. Y. et al. Sensitive tumour detection and classification using plasma cell-free DNA methylomes. Nature 563, 579–583 (2018).

[4]. Capper, D. et al. Practical implementation of DNA methylation and copy-number-based CNS tumor diagnostics: the Heidelberg experience. Acta Neuropathol. 136, 181–210 (2018).

[5]. Heyn, H. & Esteller, M. DNA methylation profifiling in the clinic: applications and challenges. Nat. Rev. Genet. 13, 679–692 (2012).

[6]. Sturm, D. et al. New brain tumor entities emerge from molecular classification of CNS-PNETs. Cell 164, 1060–1072 (2016).

[7]. Sharma, T. et al. Second-generation molecular subgrouping of medulloblastoma: an international meta

analysis of Group 3 and Group 4 subtypes. Acta Neuropathol. 138, 309–326 (2019).

[8]. Meissner, A. Epigenetic modifications in pluripotent and differentiated cells. Nat. Biotechnol. 28, 1079–1088 (2010).

[9]. Varley, K. E. et al. Dynamic DNA methylation across diverse human cell lines and tissues. Genome Res. 23, 555–567 (2013).

[10]. Jones, P. A. & Baylin, S. B. The epigenomics of cancer. Cell 128, 683–692 (2007).

[11]. Kelly, T. K., De Carvalho, D. D. & Jones, P. A. Epigenetic modifications as therapeutic targets. Nat. Bio technol. 28, 1069–1078 (2010).

[12]. Schwarzenbach, H., Hoon, D. S. B. & Pantel, K. Cell-free nucleic acids as biomarkers in cancer patients. Nat. Rev. Cancer 11, 426–437 (2011).

[13]. Corcoran, R. B. & Chabner, B. A. Application of cell-free DNA analysis to cancer treatment. N. Engl. J. Med. 379, 1754–1765 (2018).

[14]. Diaz, L. A. & Bardelli, A. Liquid biopsies: genotyping circulating tumor DNA. J. Clin. Oncol. 32, 579–586 (2014).

[15]. Wan, J. C. M. et al. Liquid biopsies come of age: towards implementation of circulating tumor DNA. Nat. Rev. Cancer 17, 223–238 (2017).

[16]. Brennan, C. W. et al. Te somatic genomic landscape of glioblastoma. *Cell* **155, 462–477 (2013).**

[17]. Capper, D. et al. DNA methylation–based classifcation of central nervous system tumours. *Nature* **555, 469–474 (2018).**

[18]. Ceccarelli, M. et al. Molecular profling reveals biologically discrete subsets and pathways of progression in difuse glioma. *Cell* 164**, 550–563 (2016).**

[19]. Sturm, D. et al. Hotspot mutations in *H3F3A* **and** *IDH1* **defne distinct epigenetic and biological subgroups of glioblastoma.** *Cancer Cell* 22**, 425–437 (2012).**

[20]. Varley, K. E. et al. Dynamic DNA methylation across diverse human cell lines and tissues. Genome Res. 23, 555–567 (2013).

[21]. Rodríguez-Paredes, M. & Esteller, M. Cancer epigenetics reaches mainstream oncology. Nat. Med. 17, 330–339 (2011).

[22]. Fernandez, A. F. et al. A DNA methylation fingerprint of 1628 human samples. Genome Res. 22, 407–419 (2012).

[23]. Capper, D. et al. DNA methylation-based classification of central nervous system tumours. Nature 555, 469–474 (2018).

[24]. Wiestler, B. et al. Assessing CpG island methylator phenotype, 1p/19q codeletion, and MGMT promoter methylation from epigenome-wide data in the biomarker cohort of the NOA-04 trial. Neuro Oncol. 16, 1630–1638 (2014).

[25]. Hovestadt V et al. Robust molecular subgrouping and copy-number profiling of medulloblastoma from small amounts of archival tumour material using high-density DNA methylation arrays. Acta Neuropathol. 125, 913–916, doi:10.1007/s00401-013-1126-5 (2013).

[26]. Capper, D. et al. DNA methylation-based classifification of central nervous system tumours. Nature 555, 469–474 (2018).

[27]. Krijthe, J. H. Rtsne: T-distributed stochastic neighbor embedding using Barnes-Hut implementation. R package version 0.15, https://cran.r-project.org/web/packages/Rtsne/index.html (2015).

[28]. Maaten, Lvd & Hinton, G. Visualizing data using t-SNE. J. Mach. Learn. Res. 9, 2579–2605 (2008).

[29] Zhao H, Li Y, Wang J. A Convolutional Neural Network and Graph Convolutional Network Based Method for Predicting the Classification of Anatomical Therapeutic Chemicals. Bioinformatics (2021).

[30] Liu Q, Hu Z, Jiang R, Zhou M. DeepCDR: a hybrid graph convolutional network for predicting cancer drug response. Bioinformatics (2020).

[31] Capper, D., Jones, D., Sill, M., Hovestadt, V., Schrimpf, D., Sturm, D., Koelsche, C., Sahm, F., Chavez, L., Reuss, D. E., Kratz, A., Wefers, A. K., Huang, K., et al. DNA methylation-based classification of central nervous system tumours. Nature, 555(7697), 469–474 (2018).

[32] Sun, G., Zhang, X., Liang, J., Pan, X., Zhu, S., Liu, Z., Armstrong, C. M., Chen, J., Lin, W., Liao, B., Lin, T., Huang, R., et al. Integrated Molecular Characterization of Fumarate Hydratase-deficient Renal Cell Carcinoma. Clinical cancer research : an official journal of the American Association for Cancer Research, 27(6), 1734–1743 (2021).

[33] Goldmann, T., Schmitt, B., Müller, J., Kröger, M., Scheufele, S., Marwitz, S., Nitschkowski, D., Schneider, M. A., Meister, M., Muley, T., Thomas, M., Kugler, C., Rabe, K. F., Siebert, R., Reck, M., & Ammerpohl, O. DNA methylation profiles of bronchoscopic biopsies for the diagnosis of lung cancer. Clinical epigenetics, 13(1), 38 (2021).

[34] Cristiano, S., Leal, A., Phallen, J., Fiksel, J., Adleff, V., Bruhm, D. C., Jensen, S. Ø., Medina, J. E., Hruban, C., White, J. R., Palsgrove, D. N., Niknafs, N., et al. Genome-wide cell-free DNA fragmentation in patients with cancer. Nature, 570(7761), 385–389 (2019).

[35] Melnikov, A.A., Scholtens, D., Talamonti, M.S., Bentrem, D.J., and Levenson, V.V. Methylation profifile of circulating plasma DNA in patients with pancreatic cancer. J. Surg. Oncol. 99, 119–122 (2009).

[36] Shen, S.Y., Singhania, R., Fehringer, G., Chakravarthy, A., Roehrl, M.H.A., Chadwick, D., Zuzarte, P.C., Borgida, A., Wang, T.T., Li, T., et al. Sensitive tumour detection and classifification using plasma cell-free DNA methylomes. Nature 563, 579–583 (2018).

[37] Sina, A.A.I., Carrascosa, L.G., Liang, Z., Grewal, Y.S., Wardiana, A., Shiddiky, M.J.A., Gardiner, R.A., Samaratunga, H., Gandhi, M.K., Scott, R.J., et al. Epigenetically reprogrammed methylation landscape drives the DNA self-assembly and serves as a universal cancer biomarker. Nat. Commun. 9, 4915 (2018).

[38] Jiang, P., Sun, K., Tong, Y.K., Cheng, S.H., Cheng, T.H.T., Heung, M.M.S., Wong, J., Wong, V.W.S., Chan, H.L.Y., Chan, K.C.A., et al. Preferred end coordinates and somatic variants as signatures of circulating tumor DNA associated with hepatocellular carcinoma. Proc. Natl. Acad. Sci. U S A 115, E10925–E10933 (2018).

[39] Mouliere, F., Chandrananda, D., Piskorz, A.M., Moore, E.K., Morris, J., Ahlborn, L.B., Mair, R., Goranova, T., Marass, F., Heider, K., et al. Enhanced detection of circulating tumor DNA by fragment size analysis. Sci. Transl. Med. 10, eaat4921 (2018a).

[40] Snyder, M.W., Kircher, M., Hill, A.J., Daza, R.M., and Shendure, J. Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of origin. Cell 164, 57–68 (2016).

[41] Ulz, P., Thallinger, G.G., Auer, M., Graf, R., Kashofer, K., Jahn, S.W., Abete, L., Pristauz, G., Petru, E., Geigl, J.B., et al. Inferring expressed genes by whole-genome sequencing of plasma DNA. Nat. Genet. 48, 1273–1278 (2016).

Acknowledgements

D-Q. W. and E. W. conceived the research, X. J. and Z. L. designed and conducted the research, H. W., Y. C., Q. W., X. M., M. J., B. Z. and X. Y. collected the data, and checked the program, X. J. and Z. L. wrote the manuscript. All authors reviewed the manuscript.

Author contributions

This work is supported in part by funds from the National Science Foundation of China (Grant No. 32070662, 61832019, 32030063), the Key Research Area Grant 2016YFA0501703 of the Ministry of Science and Technology of China, the Science and Technology Commission of Shanghai Municipality (Grant No. 19430750600), as well as SJTU JiRLMDS Joint Research Fund and Joint Research Funds for Medical and Engineering and Scientific Research at Shanghai Jiao Tong University (YG2021ZD02).

Competing interests

The authors declare no competing interest.

Data availability

The complete methylation data required for the construction and training of the classifier have been deposited in TCGA. The external validation data can be downloaded from the NCBI Gene Expression Omnibus (GEO), under accession number GSE90496, GSE155207, GSE158075, GSE164988, and GSE157272.

Code availability

All the codes used in this research are complied with Python programming language. The codes, data files and more detailed description are available at https://github.com/lizhiqi0506/SAGCN.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Tables.pdf
- Tables.pdf
- suppleTable1clinicalinfostatistic.xlsx
- suppleTable2selectedprobes350annotation.csv