Noninvasive Lung Cancer Early Detection via Deep Methylation Representation Learning

Xiangrui Cai[1], Jinsheng Tao[2], Shichao Wang[1], Zhiyu Wang[1], Jiaxian Wang[1], Mei Li[1], Hong Wang[2], Xixiang Tu[2], Hao Yang[2], Jian-Bing Fan[2*], Hua Ji[1*]

[1] Advanced Medical Data Research Center, CS, Nankai University, Tianjin, China
[2] AnchorDx Medical Co., Guangzhou, China
caixr@nankai.edu.cn, jinsheng_tao@anchordx.com, wangshichao@dbis.nankai.edu.cn,
wangzhiyu@mail.nankai.edu.cn, wangjiaxian@dbis.nankai.edu.cn, LIMei-666@mail.nankai.edu.cn,
hong_wang@anchordx.com, xixiang_tu@anchordx.com, hao_yang@anchordx.com,
jianbing_fan@anchordx.com, hua.ji@nankai.edu.cn

Abstract

Early detection of lung cancer is crucial for five-year survival of patients. Compared with the pathological analysis and CT scans, the circulating tumor DNA (ctDNA) methylation based approach is noninvasive and cost-effective, and thus is one of the most promising methods for early detection of lung cancer. Existing studies on ctDNA methylation data measure the methylation level of each region with a predefined metric, ignoring the positions of methylated CpG sites and methylation patterns, thus are not able to capture the early cancer signals. In this paper, we propose a blood-based lung cancer detection method, and present the first ever study to represent methylation regions by continuous vectors. Specifically, we propose DeepMeth to regard each region as a one-channel image and develop an auto-encoder model to learn its representation. For each ctDNA methylation sample, DeepMeth achieves its representation via concatenating the region vectors. We evaluate DeepMeth on a multicenter clinical dataset collected from 14 hospitals. The experiments show that DeepMeth achieves about 5%-8% improvements compared with the baselines in terms of Area Under the Curve (AUC). Moreover, the experiments also demonstrate that DeepMeth can be combined with traditional scalar metrics to enhance the diagnostic power of ctDNA methylation classifiers. DeepMeth has been clinically deployed and applied to 450 patients from 94 hospitals nationally since April 2020.

Introduction

Lung cancer has been one of the deadliest cancers (Didkowska et al. 2016). The prognosis of lung cancer is highly correlated with the stage of the disease at diagnosis, which affects the five-year survival rate of patients. For instance, the five-year survival rate decreases greatly from 85% for stage IA to 6% for stage IV (Torre, Siegel, and Jemal 2016). Therefore, early detection of lung cancer is crucial for saving lives and reducing medical costs, which has significant clinical value and social impact.

DNA methylation, found primarily at CpG dinucleotides, is an epigenetic mechanism used by cells to controls gene expression (Deaton and Bird 2011). Due to its significance in the etiology of diseases, DNA methylation analysis has been a powerful tool in cancer diagnosis. Much research has been done to measure tissue-based DNA methylation level, such as Beta-value (Bibikova and Fan 2009), methylation entropy (Xie et al. 2011), and the percentage of comethylation (Liang et al. 2019).

Compared to the tissue-based DNA methylation, the way to obtain circulating tumor DNA (ctDNA) methylation in blood is more noninvasive and cost-effective, which makes it more valuable for early detection of cancers. Recent research has demonstrated that ctDNA methylation is exquisitely specific for lung cancer detection (Guo et al. 2017; Liang et al. 2021). Therefore, to analyze ctDNA methylation is one of the most promising way for early detection of lung cancer.

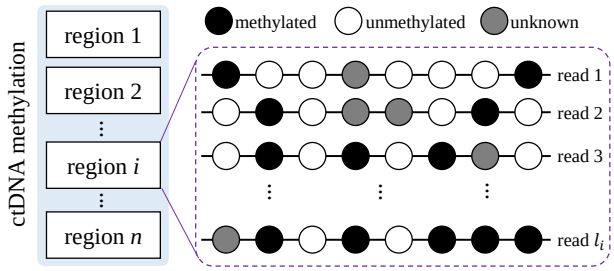

Figure 1: Illustration of a ctDNA methylation sample. Each sample usually contains thousands of regions, and each region consists of tens of reads. The CpG sites in a read has three status, i.e., methylated (black), unmethylated (white) and unknown (grey).

Figure 1 illustrates the raw data of a ctDNA methylation sample obtained by high throughput DNA bisulfite sequencing. It consists of thousands of regions (mostly CpG islands), which are the key genomic fragments identified by biomedical scientists (Liang et al. 2019). Each region contains tens of reads. Each read is comprised of CpG sites. There are three methylation status for a CpG site, i.e., methylated, unmethylated, and unknown. As observed from Figure 1, the raw data of ctDNA methylation samples contains various methylation patterns and large amounts of information,

*Corresponding Authors.

which makes ctDNA methylation analysis suffer from the curse of dimensionality.

Desipte some works on the metrics of the methylation level, there are mainly two reasons that prevent them from applications on ctDNA methylation analysis for lung cancer. First, most of them are designed for tissue methylation samples, where malignant signals are stronger and thus easier to be detected compared to ctDNA methylation samples. Second, existing metrics typically represent each region by a scalar, while ignoring positions of methylated CpG sites and methylation patterns, which are predictive features for cancer detection (Affinito et al. 2020; Gatev et al. 2020).

In this paper, we propose DeepMeth, a novel method to analyze ctDNA methylation data for lung cancer early detection. To our best knowledge, it is the first study that analyzes the raw data of ctDNA methylation with deep learning. Specifically, DeepMeth first learns the continuous representations of regions with a residual-based auto-encoder. Then, it obtains the representation of each ctDNA methylation sample by concatenating the region vectors. Finally, DeepMeth trains a classifier on the ctDNA methylation data to detect lung cancer. To address the curse of dimensionality problem, each region is regarded as an independent sample for the auto-encoder according to the Occam's Razor [1] (Schaffer 2015). The reason behind is that the correlations between regions are limited within a methylation sample (Eckhardt et al. 2006). We evaluate DeepMeth on a clinical dataset LC-Meth. The dataset contains the ground truth (i.e., benign or malignant) from the pathological analysis of lung nodules. Compared to the baselines, DeepMeth achieves 5%-8% improvements on LC-Meth in terms of Area Under the Curve (AUC).

In summary, we make the following main contributions:

- In this paper, we propose a blood-based method, DeepMeth, for early prediction of lung cancer. Compared with pathological analysis and CT scans, the proposed method is noninvasive and cost-effective. Since April 2020, DeepMeth has been clinically applied to 450 patients from 94 hospitals.
- To the best of our knowledge, it is the first study that represents methylation regions by continuous vectors. To avoid the curse of dimensionality, it regards each methylation region independently and develops an auto-encoder to learn the rich semantics (e.g., positions of methylated CpG sites) hidden in region layouts.
- We conduct extensive experiments on a clinical dataset collected from 14 hospitals, LC-Meth. The experiments show that DeepMeth outperforms four state-of-the-art baselines by 5%-8% in terms of AUC. In addition, it has also been demonstrated that our auto-encoder features can be combined with traditional metrics to enhance prediction performance.

Related Work

In this section, we first review the current available metrics that measure the methylation level of a region. Then, we introduce recent deep learning-based studies on DNA methylation data.

Metrics of Methylation Level

Much research has been done to measure DNA methylation level. Beta-value (Bibikova and Fan 2009) has been the most widely used metric to measure the methylation level of a region. It calculated the ratio of the methylated probe intensity, and was utilized as the input for cancer detection (Li et al. 2018; Wang and Wang 2018; Levy et al. 2020). In order to satisfy the range of Gaussian distribution, some works proposed transformations of Beta-value, e.g., M-value (Du et al. 2010), but degraded the performance.

Although widely used, Beta-value only calculated the frequency of methylated sites in a region, ignoring the co-methylation information within the region (i.e., consecutive methylated sites). To address the problem, several metrics have been proposed. Specifically, (Xie et al. 2011) proposed to utilize the entropy concept in information theory to quantify the methylation level of a region. (Landan et al. 2012) proposed the Epipolymorphism metric that measured the methylation level of the set of four CpGs by considering the frequency of epi-alleles in the population. (Guo et al. 2017) identified 771 million methylation haplotype blocks (Shoemaker et al. 2010) based on 61 Whole-Genome Bisulfite Sequencing (WGBS) data from human primary tissues. Then, they defined Methylation Haplotype Load (MHL) to quantify the methylation level of a block, which is the weighted average of the fraction of fully methylated haplotypes and substrings at different lengths. Recently, the percentage of co-methylation has aroused much attention from the community (Wang et al. 2016; Gomez et al. 2019; Affinito et al. 2020), and exhibited its power in the classification of cancers like the lung cancer (Liang et al. 2019) and the breast cancer (Sun et al. 2019). It first defined a read as co-methylated if there exist a methylated sites within b consecutive CpG sites ($a \leq b$ and they are hyper-parameters) (Liang et al. 2019). Then, it calculated the ratio of co-methylated reads to the overall reads within a region.

Despite the progress, the limitations of the aforementioned metrics are two folds. First, most of them were designed for tissue methylation samples, where malignant signals are stronger and easier to be detected compared with ctDNA methylation samples. Second, previous studies typically represented each region by a scalar, and might miss rich semantics conveyed by region layouts. In this paper, we propose an unsupervised learning-based method to represent each methylation region by a continuous vector, which captures more semantic information of raw methylation data.

Deep Learning for DNA Methylation

Deep learning has accelerated the research of DNA methylation data analysis. Some studies applied deep learning on DNA methylation data for cancer detection (Wang and Wang 2018; Titus, Bobak, and Christensen 2018; Levy et al. 2020; Macías-García et al. 2020) and age prediction (Galkin et al. 2021). Specifically, (Wang and Wang 2018; Levy et al. 2020) constructed a deep learning model to detect lung cancers, and fed Beta-value into the model as the input. (Ti-

[1] https://en.wikipedia.org/wiki/Occam's_razor

tus, Bobak, and Christensen 2018) and (Macías-García et al. 2020) performed feature engineering and exploited auto-encoder for breast cancer and breast cancer recurrence diagnosis respectively. Desipte the progress, these methods mainly focused on the tissue data from The Cancer Genome Atlas (TCGA) data portal, and degraded the performance when directly applied on ctDNA methylation data. On the other hand, existing studies usually exploited metrics (e.g., Beta-value) or feature engineering to analyze methylation data, which required experiences of biological experts and might introduce bias.

In addition, some studies improve the usability of methylation data with deep learning. These studies typically imputed the status of unknown CpG sites to improve the analysis of methylation data. Specifically, DeepCpG (Angermueller et al. 2017; Ni et al. 2019; Tian et al. 2019) proposed to employ convolutional neural networks to predict the unknown methylation status after bisulfite sequencing. Since the scales of DNA methylation datasets are usually very small, MethCancer-Gen (Choi and Chae 2020) employed a variational auto-encoder (Kingma and Welling 2013) model to generate synthesis methylation data.

In this paper, we propose DeepMeth to model raw methylation data and capture rich semantics hidden in the region layouts. In addition, instead of tissue-based methods, this paper developed a blood-based method to detect lung cancers, which is non-invasive and cost-effective.

Methodology

In this section, we first introduce the problem definition. Then, we describe the proposed DeepMeth in detail.

Problem Definition

According to the microarray-based sequencing technologies (e.g., the Illumina HumanMethylation450 (450K) array), a ctDNA methylation sample m consists of a sequence of n regions, i.e., $m = [R_1, R_2, \ldots, R_n]$, where R_i refers to the i-th region. A region R_i contains tens of reads and can be formulated as:

$$R_i = \{r^i(1), r^i(2), \ldots, r^i(k_i)\}\,,$$

where $r^i(j)$ is the j-th read in R_i and k_i denotes the number of reads within the region. Notice that the methylation reads from the same region have the same length, while the length of reads from different regions could be different. Each read is a sequence of CpG sites, and each CpG site has a methylation status, which is denoted by methylated (C), unmethylated (T), or unknown(N). The read $r^i(j)$ with l_j sites can be formulated as follows:

$$r^i(j) = [s^i(j)_0, s^i(j)_1, \ldots, s^i(j)_{l_j}]\,,$$
$$s^i(j)_k \in \{C, T, N\}, k \in \{1, 2, \ldots, l_j\}\,.$$

Given a ctDNA methylation sample m, the goal of lung cancer detection is to predict whether the input is benign or malignant. Specifically, we divide the classification into two phases. In the first phase, a representation learning model Φ is applied to learn methylation representations. In the second phase, we use a classifier $\mathcal{F}$ to predict the pathology class of m. Formally,

$$\boldsymbol{v}_m = \Phi(m) = \Phi([R_1, R_2, \ldots, R_n])\,,$$
$$y_m = \mathcal{F}(\boldsymbol{v}_m) \in \{\text{Benign}, \text{Malignant}\}\,,$$

where $\boldsymbol{v}_m$ refers to the methylation representation of m, Φ the representation learning model, and $\mathcal{F}$ the classifier.

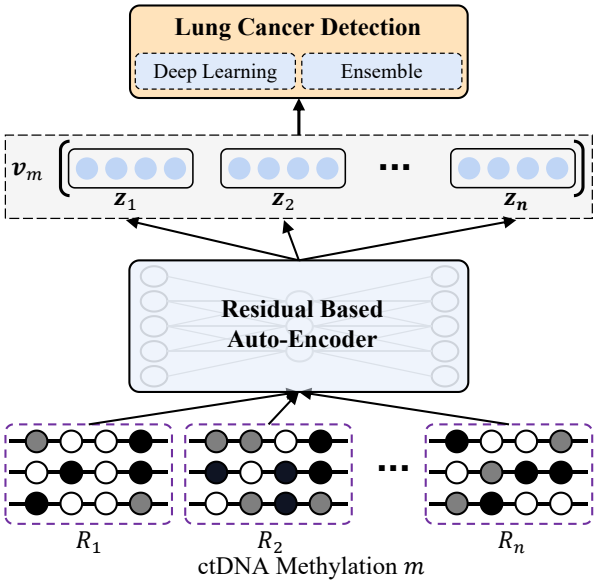

Figure 2: The framework of DeepMeth. It can be divided into two phases. The first phase is region representation learning, aiming to extract biomedical features from the methylation regions. The second phase is classification. Several classifiers are employed to classify the methylation into malignant or benign pathology classes.

DeepMeth

Figure 2 shows the overall framework of DeepMeth. DeepMeth contains two phases. Given a ctDNA methylation m, we regard the regions within m independently of each other. We develop a residual-based auto-encoder model to learn region representations in the first phase. Then in the second phase, we employ a widely used classifier for pathology classification based on the sequence of region vectors. Next we introduce the details of DeepMeth.

Auto-Encoder for Region Representations DeepMeth, for the first time, represents each methylation region by a continuous vector. To capture more semantic information of the methylation region, we develop an auto-encoder model to the learn the methylation region representation. As described in the Introduction, the number of ctDNA methylation samples is small due to the expensive cost for collection. However, the number of regions and that of reads are very large. Thus, analyzing ctDNA methylation suffers from the curse of dimensionality problem. Existing study (Eckhardt et al. 2006) has found that the correlations among regions are very weak. Therefore, DeepMeth regards each region

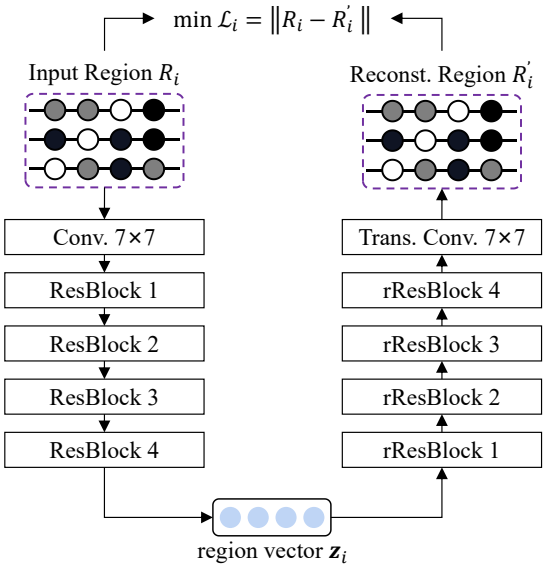

Figure 3: The architecture of the residual-based auto-encoder. We treat the regions independently and take them as the input for the auto-encoder.

independently of each other and learns a vector representation of a region. The region representation model has a large number of regions as the training data. Specifically, the number of training instances increase from #samples to #regions × #samples. We develop a residual-based auto-encoder model to learn region representations. The overall architecture of the residual-based auto-encoder is depicted in Figure 3. Given an input region R_i, the encoder compress the raw data in to a region $\boldsymbol{z}_i$ through a sequence of neural layers. Then the decoder reconstruct the region R_i' with a sequence of reversed operations. we detail the encoder, the decoder, and the training objective in the auto-encoder next.

We employ the ResNet (He et al. 2016) as the backbone of DeepMeth encoder. Formally,

$$\boldsymbol{z}_i = \text{ResNet}(R_i)\,,$$

where $\boldsymbol{z}_i \in \mathbb{R}^h$ refers to the learned region vector of the i-th region R_i.

Specifically, The encoder starts with a down-sample operation, and followed by four residual blocks (ResBlock). The input down-sample operation contains a convolutional layer with 7×7 filters, a batch normalization (BN) (Ioffe and Szegedy 2015) layer, a rectified linear unit (ReLU) (Glorot, Bordes, and Bengio 2011) activation layer, and a max-pooling layer with 3×3 filters (Figure 4). The large kernel convolution layer can capture the local region biological methylation semantics. The ResBlock is consisted of two convolution blocks. Each block contains a 3×3 convolution layer, a batch normalization, a rectified linear unit, and a residual connection (Figure 3). The residual connections are able to migrate the biological features from different levels and alleviate the information loss problem happened in deep neural networks. Table 1 lists the configurations of the encoder layers in detail. Finally, we employ a linear pooling

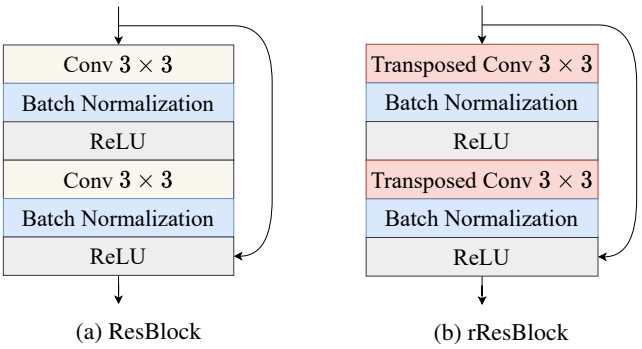

Figure 4: Detail structures of ResBlock and rResBlock.

Block	Encoder		Block	Decoder	
	input	output		input	output
ResBlock1	64	64	rResBlock1	64	64
ResBlock2	64	128	rResBlock2	128	64
ResBlock3	128	256	rResBlock3	256	128
ResBlock4	256	512	rResBlock4	512	256

Table 1: Block configurations in the auto-encoder. The input and the output fields indicate the sizes of input channel and output channel of each block respectively.

layer to get the encoded region vector.

As for the decoder, we introduce the reversed ResNet (rResNet) to reconstruct the regions from the region vectors. Formally, we denote the reconstructed region by R', which is obtained by the reversed ResNet (rResNet).

$$R_i' = \text{rResNet}(\boldsymbol{z}_i)\,,$$

The rResNet is designed following three principles: (1) following reversed data flow; (2) replacing the convolution layer with the transposed convolution layer (Dumoulin and Visin 2016); (3) preserving the residual connection. Following the principles, we also build the reversed residual block (rResBlock) corresponding to the ResBlock in the encoder (Figure 4). Table 1 also shows the detailed rResBlock.

Similar to conventional auto-encoder models, we take Mean Square Error (MSE) as our region reconstruction criterion:

$$\mathcal{L} = \frac{1}{n \times N} \sum_{i=1}^{n \times N} ||R_i' - R_i||_2^2\,,$$

where N is the number of ctDNA methylation samples in the training set.

Lung Cancer Detection To detect the methylation signal from the auto-encoded region vectors, we study two strategies. The first one is to apply deep learning models to the region vectors directly. We use multi-layer perceptrons (MLP), convolution neural network (CNN), and the recurrent neural network (RNN). MLP is a widely used deep learning classifier, which learn the relationships between methylation features and the pathology class by changing the weights of its neurons. The non-linear activation layer

between the perceptrons enable learning the non-linear relationships between the pathology classes and the classification features. However, MLP is prone to overfit due to the massive training parameters. CNN is usually consisted of multiple convolution layers and a fully connection layer. The convolutional layers share parameters across kernels (local receptive fields), thus reduce the number of parameters to train and have a better generalization ability than MLP. RNN is designed for sequential data, which is suitable for modeling the sequence of methylation region vectors.

On the other hand, we build traditional machine learning classification models based on the auto-encoded region vectors. In this case, we need to combine them together first to obtain the ctDNA methylation representation. We use concatenation and mean as the aggregation methods, namely,

$$\boldsymbol{v}_m = \begin{cases} [\boldsymbol{z}_0; \boldsymbol{z}_i; \ldots, \boldsymbol{z}_n], & \text{(concatenation)} \\ \frac{1}{n}\sum_{i=0}^{n}(\boldsymbol{z}_i), & \text{(mean)} \end{cases},$$

where $\boldsymbol{z}_i$ is the hidden vector encoded from i-th region. $\boldsymbol{v}_m$ is the classification feature.

Following previous studies (Guo et al. 2017; Liang et al. 2019, 2021), we employ three widely used ensemble classifiers, namely, Random Forest, XGBoost (Chen and Guestrin 2016), and LightGBM (Ke et al. 2017). The Random Forest is an early and widely used classical tree-based ensemble learning model for classification. It creates many decision trees from the subset of the problem, thus it overcomes the problem of overfitting within one decision tree. The class selected by most decision trees is returned as the output of the random forest. XGBoost is an optimized distributed gradient boosting framework. It leverages a weighted quantile sketch for approximate tree learning methods. LightGBM is another distributed gradient boosting framework for machine learning. The LightGBM has better scalability than the XGBoost by leveraging the gradient-based one-side sampling and the exclusive feature bundling techniques.

We conduct extensive experiments to compare the classification models. We notice that the ensemble methods, i.e., Random Forest, XGBoost, and LightGBM, are able to achieve high performance without careful tuning.

Dataset

We evaluate DeepMeth on a clinical dataset, LC-Meth. It contains 424 patients in total and their pathological analysis results. The dataset were collected by Company X from the thoracic department of 14 hospitals. The criteria for selecting subjects are as follows: adult patients who are no younger than 18 years old; both male and female; the size of a single nodule is between 5 and 30 millimeters detected by standard or Low-Dose CT screening; nodule types include solid, part-solid, and pure ground-glass. LC-Meth also excluded several subjects, such as pregnant or lactating females, patients with metastasis symptoms, patients without confirmed pathological diagnosis after surgery, and patients with confirmed cancer 2 years prior to enrollment. LC-Meth finally collected pathological data information of 424 plasma samples. For each sample, it contains 10180 preselected lung cancer–specific methylation regions. Each region is composed of multiple Illumina sequencing reads (Sandoval et al. 2011). LC-Meth contains all the information of the pathological examination, and the detailed deidentified clinical information (including demographics, LDCT imaging reports, and pathology reports) of the subjects. Therefore, the dataset contains both the methylation data and the ground truth of benign or malignant.

Category		**Train.**		**Valid.**		**Test**	
		M	B	M	B	M	B
Age	0-40	6	16	5	6	2	2
	41-55	76	52	19	17	9	6
	56-70	96	31	26	7	14	7
	≥ 71	14	0	4	0	7	2
Gender	Male	39	46	15	11	5	8
	Female	58	36	13	13	12	3
Nodule	Nonsolid	37	13	11	3	7	2
	Partially Solid	37	28	10	6	6	4
	Solid	23	41	7	15	2	7
Total Number		192	106	54	30	27	15

Table 2: Statistics in training, validation and test sets stratified by age group, gender and nodule type. (M: Malignant, B: Benign).

The blood samples of the enrolled patients were collected in Streck cell-free DNA BCT tubes (Streck, catalog 218962) and were shipped to Company X's certified molecular diagnosis laboratory. Then, we separated plasma from the blood samples immediately and stored it at –80°C. We performed Bisulfite conversion using the EZ DNA Methylation-Lightning Kit (catalog D5031, Zymo Research). We conducted targeted genome methylation analysis on 10 ng input ctDNA, and used a custom lung cancer methylation specific panel, which consists of 10180 preselected regions.

We randomly split the training, validation, and test sets by 7:1:2, and made the distribution of both the demographics of the patients and the morphological features of the nodules in accordance with the proportion. Finally, we obtained 298 training samples, 42 test samples, and 84 validation samples. We run the experiments on 10 random splits and reported the average performance. We present the statistics of the training, test, and validation sets in Table 2.

Experiments

In this section, we describe the details of the experimental evaluation. We first introduce the baselines and the implementation details. Then, we vary the size of representation vectors and present the results of region representation. Given the same representation, we report the results of two aggregation methods, and compare DeepMeth with four baseline metrics to evaluate the performance of DeepMeth. Moreover, we also show the results of experiments that combine DeepMeth with traditional metrics.

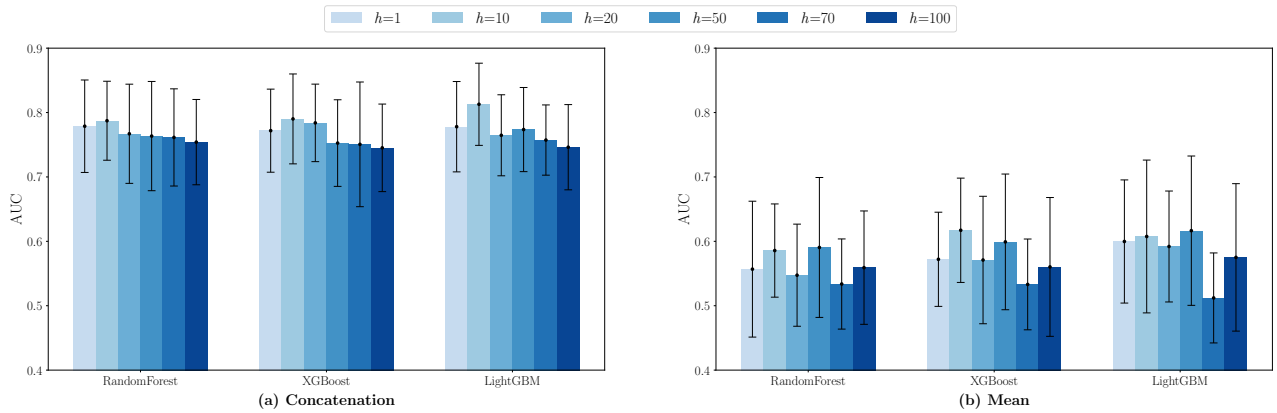

Figure 5: The influence of the size of region vectors. (a) ctDNA Methylation is represented by concatenation of region vectors; (b) ctDNA Methylation is represented by mean of region vectors. We varies the size of region vectors from 1 to 100. Each bar group shows the AUC scores under different classifiers.

Baselines

We compare DeepMeth with four widely-used cancer detection methods, which utilize the traditional metrics of the methylation level. Specifically, the metrics include the Percentage of Co-Methylation (PCM) (Liang et al. 2019), the Methylation Frequency (MF) (Bibikova and Fan 2009), Methylation Entropy (ME) (Xie et al. 2011), and Methylation Haplotype Load (MHL) (Guo et al. 2017). Notice that MF is the same with Beta-value (the details of these metrics can be seen in the Appendix). Then, the baseline methods employ traditional machine learning classifiers (e.g., Random Forest) to predict the diagnosis of inputs.

Implementation Details

We implement the auto-encoder of DeepMeth with PyTorch (Paszke et al. 2019). We use Adam (Kingma and Ba 2015) to optimize the parameters. The learning rate is set to 1e-4 and the weight-decay is 1e-5. The training batch size is 32.

To compare with existing methods, we implement three ensemble classifiers on all baseline metrics and our region vectors. Specifically, we implement RandomForest, XGBoost, and LightGBM with scikit-learn (Pedregosa et al. 2011). We also perform grid search on the validation set to choose the best hyper-parameters (listed in the Appendix).

We train the auto-encoder on with a Nvidia RTX 2080Ti GPU. All models are run on the same random splits. We report the mean AUC scores and their variations. We also present the receiver operating characteristic (ROC) curves in the Appendix. The source code is available at https://github.com/XiangruiCAI/DeepMeth/.

Influence of the Size of Region Vectors

In order to analyze the size of region vectors, we conduct the experiments using various sizes of region vectors, including 1, 10, 20, 50, 70, and 100. The learned vectors are processed to form a methylation vector with two methods, i.e., mean and concatenation. Then, based on the methylation vector, we employ three ensemble classification models (i.e., Random Forest, XGBoost, and LightGBM) for cancer detection.

The results are reported in Figure 5(a). We report the mean and variance of the AUC scores on the test set of 10 random splits. When the dimensionality is set to 1, a region is represented by a scalar, which is similar to the traditional metrics. As we can observe, setting the dimensionality to 1 is ineffective to encode the information of methylation patterns within a region, resulting in the sub-optimal performance. As the size of region vectors increases, the performance of the three classifiers increases first and then decreases. We can observe that the classifiers perform best when the dimensionality is set to 10. On the other hand, in Figure 5(b), we compare two aggregation methods, i.e., mean and concatenation. It can be observed that concatenation achieves much better performance than mean. The mean aggregation behaves like a random guess. The reason is that a small number of discriminative region vectors are eliminated in the mean aggregation due to the large number of regions. The tumor signals are drowned out through the mean aggregation.

Comparison of DeepMeth and SOTA Methods

We compare DeepMeth with four baseline metrics on LC-Meth. DeepMeth is configured with two sizes of region vectors, i.e., dim=1 and dim=10. The aggregation method is concatenation. We perform prognosis classification with three ensemble classifiers, i.e., RandomForest, XGBoost, and LightGBM. As shown in Table 3, ME performs the best among the baselines. Though we set the size of region vectors to 1, DeepMeth outperforms ME by 3% to 5.6% across three classifiers in terms of AUC. This extreme setting indicates that the scalar feature learned by the auto-encoder is better than the hand-craft metrics. When the size of region vector is set to 10, DeepMeth achieves the best performance, i.e., 5% to 8% higher than ME in terms of AUC. We can observe that PCM (3-5) performs better than PCM (2-3). One possible explanation is that PCM (3-5) has a more relaxed conditions on co-methylation. We also notice that MHL has poor performance on the LC-Meth dataset. The identification of methylation haplotype blocks (MHB) is the premise for calculating the MHL score. It heavily depends on the tis-

Region Representations	Classifiers		
	RandomForest	XGBoost	LightGBM
PCM (2-3)	0.6692±0.0675	0.6097±0.0943	0.6055±0.1458
PCM (3-5)	0.7074±0.0666	0.6606±0.1132	0.6634±0.1006
ME	0.7470±0.0818	0.7055±0.0901	0.7415±0.0800
MF	0.7066±0.0905	0.6281±0.0880	0.7057±0.0859
MHL($w_i = i$)	0.7004±0.0945	0.6110±0.1103	0.6834±0.1010
DeepMeth ($h = 1$)	0.7770±0.0718	0.7712±0.0645	0.7770 ±0.0702
DeepMeth ($h = 10$)	**0.7860±0.0613**	**0.7892±0.0698**	**0.8117±0.0638**

Table 3: Comparsion of DeepMeth and the baselines. DeepMeth achieves best performance on LC-Meth in terms of AUC. Compared to the best baselines (ME), DeepMeth improves the AUC by about 5%-8% across the three classifers.

Method	RandomForest	XGBoost	LightGBM
DeepMeth($h = 10$)	0.7860±0.0613	0.7892±0.0698	0.8117±0.0638
DeepMeth($h = 10$) **+ ME**	**0.8026±0.0703**	**0.8195±0.0694**	**0.8326±0.0527**
DeepMeth($h = 10$) **+ MF**	0.8017±0.0627	0.8114±0.0580	0.8309±0.0592
DeepMeth($h = 10$) **+ MHL**	0.7919±0.0643	0.8114±0.0630	0.8272±0.0578
DeepMeth($h = 10$) **+ PCM(2-3)**	0.7930±0.0661	0.8032±0.0671	0.8262±0.0570
DeepMeth($h = 10$) **+ PCM(3-5)**	0.7928±0.0633	0.8030±0.0707	0.8200±0.0599

Table 4: Representing methylation regions with combined features. The size of region vectors is set to be 10. Combining DeepMeth and a traditional metric is able to achive higher AUC scores than DeepMeth only.

sue data, which degrades the performance of MHL on new datasets. In addition, it can be observed that the LightGBM performs the best among the three the ensemble classifiers.

Combined Region Representations

To investigate the effectiveness of our auto-encoder features, we also conduct experiments that combine DeepMeth and the scalar metrics. For each region, we concatenate the DeepMeth features with the metrics to get a combined representation. We set the size of region vectors to 10. The results are reported in Table 4. As we can observe, it achieves a small improvement when combining DeepMeth with traditional metrics. It indicates that the two types of features can serve as the supplements of each other. We can also observe that LightGBM over the region vectors and ME performs the best, gaining more than 2% improvements compared with DeepMeth-only.

Comparison to Deep Learning Classifiers

We also compare DeepMeth equipped with different classifiers. Specifically, we feed the learned methylation representations into six classifiers for early detection of lung cancer. Among these classifiers, three of them are ensemble models (i.e., Random Forest, XGBoost, and LightGBM), and the others are deep learning-based models (i.e., MLP, CNN, and RNN). We report the details of the models in the Appendix.

The experimental results are shown in Table 5. It can be observed that with the ensemble classifiers, DeepMeth achieves higher AUC in lung cancer detection than with the deep learning models. One possible explanation is that the methylation dataset has 10180 regions, while the number of samples is only 424. The deep learning models have too many parameters to tune and are prone to overfitting.

Classifiers	AUC
RandomForest	0.7860 ± 0.0613
XGBoost	0.7892 ± 0.0698
LightGBM	**0.8117± 0.0638**
MLP	0.7498 ± 0.0659
CNN	0.7604 ± 0.0642
RNN	0.7027 ± 0.0693

Table 5: The comparison between DeepMeth and other classifiers. We can observe that the LightGBM achieves the best AUC and outperforms the second (XGBoost) by about 2%.

Conclusion

This paper proposes a noninvasive and cost-effective method, DeepMeth, for early detection of lung cancer. Instead of predefined metrics, we present an auto-encoder to learn the rich semantics hidden in the layouts of ctDNA methylation regions. To avoid the curse of dimensionality, we regard each methylation region independently and concatenate region vectors to represent a ctDNA methylation sample. Extensive experiments on a multicenter clinical dataset have demonstrated the effectiveness of DeepMeth compared with four state-of-the-art baselines. Moreover, our experiments also show that the auto-encoder features can be well combined with traditional metrics to enhance the prediction performance. DeepMeth has been clinically deployed national-wide in more than 94 hospitals. Our future research plan is to apply DeepMeth to the prognosis of other cancers such as liver and breast cancers.

Acknowledgments

We are grateful for so many helps from Dr. Zeyu Jiang, Dr. Zhiwei Chen, Bo Wang and Dr. Ying Zhang during the past three years. We also want to thank all of our colleagues at Nankai-AnchorDx Advanced Medical Data Research Center and Trusted AI System Laboratory.

References

Affinito, O.; Palumbo, D.; Fierro, A.; Cuomo, M.; De Riso, G.; Monticelli, A.; Miele, G.; Chiariotti, L.; and Cocozza, S. 2020. Nucleotide distance influences co-methylation between nearby CpG sites. *Genomics*, 112(1): 144–150.

Angermueller, C.; Lee, H. J.; Reik, W.; and Stegle, O. 2017. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome biology*, 18(1): 1–13.

Bibikova, M.; and Fan, J.-B. 2009. GoldenGate® assay for DNA methylation profiling. In *DNA Methylation*, 149–163. Springer.

Chen, T.; and Guestrin, C. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, 785–794. New York, NY, USA: ACM. ISBN 978-1-4503-4232-2.

Choi, J.; and Chae, H. 2020. methCancer-gen: a DNA methylome dataset generator for user-specified cancer type based on conditional variational autoencoder. *BMC bioinformatics*, 21: 1–10.

Deaton, A. M.; and Bird, A. 2011. CpG islands and the regulation of transcription. *Genes & development*, 25(10): 1010–1022.

Didkowska, J.; Wojciechowska, U.; Mańczuk, M.; and Łobaszewski, J. 2016. Lung cancer epidemiology: contemporary and future challenges worldwide. *Annals of translational medicine*, 4(8).

Du, P.; Zhang, X.; Huang, C.-C.; Jafari, N.; Kibbe, W. A.; Hou, L.; and Lin, S. M. 2010. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC bioinformatics*, 11(1): 1–9.

Dumoulin, V.; and Visin, F. 2016. A Guide to Convolution Arithmetic for Deep Learning. *ArXiv*.

Eckhardt, F.; Lewin, J.; Cortese, R.; Rakyan, V. K.; Attwood, J.; Burger, M.; Burton, J.; Cox, T. V.; Davies, R.; Down, T. A.; et al. 2006. DNA methylation profiling of human chromosomes 6, 20 and 22. *Nature genetics*, 38(12): 1378–1385.

Galkin, F.; Mamoshina, P.; Kochetov, K.; Sidorenko, D.; and Zhavoronkov, A. 2021. DeepMAge: A methylation aging clock developed with deep learning. *Aging and disease*, 12(5): 1252.

Gatev, E.; Gladish, N.; Mostafavi, S.; and Kobor, M. S. 2020. CoMeBack: DNA methylation array data analysis for co-methylated regions. *Bioinformatics*, 36(9): 2675–2683.

Glorot, X.; Bordes, A.; and Bengio, Y. 2011. Deep Sparse Rectifier Neural Networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 315–323. JMLR Workshop and Conference Proceedings.

Gomez, L.; Odom, G. J.; Young, J. I.; Martin, E. R.; Liu, L.; Chen, X.; Griswold, A. J.; Gao, Z.; Zhang, L.; and Wang, L. 2019. coMethDMR: accurate identification of co-methylated and differentially methylated regions in epigenome-wide association studies with continuous phenotypes. *Nucleic acids research*, 47(17): e98–e98.

Guo, S.; Diep, D.; Plongthongkum, N.; Fung, H.-L.; Zhang, K.; and Zhang, K. 2017. Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA. *Nature genetics*, 49(4): 635–642.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Ioffe, S.; and Szegedy, C. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *International Conference on Machine Learning*, 448–456. PMLR.

Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; and Liu, T.-Y. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Landan, G.; Cohen, N. M.; Mukamel, Z.; Bar, A.; Molchadsky, A.; Brosh, R.; Horn-Saban, S.; Zalcenstein, D. A.; Goldfinger, N.; Zundelevich, A.; et al. 2012. Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues. *Nature genetics*, 44(11): 1207–1214.

Levy, J. J.; Titus, A. J.; Petersen, C. L.; Chen, Y.; Salas, L. A.; and Christensen, B. C. 2020. MethylNet: an automated and modular deep learning approach for DNA methylation analysis. *BMC bioinformatics*, 21(1): 1–15.

Li, W.; Li, Q.; Kang, S.; Same, M.; Zhou, Y.; Sun, C.; Liu, C.-C.; Matsuoka, L.; Sher, L.; Wong, W. H.; et al. 2018. CancerDetector: ultrasensitive and non-invasive cancer detection at the resolution of individual reads using cell-free DNA methylation sequencing data. *Nucleic acids research*, 46(15): e89–e89.

Liang, W.; Chen, Z.; Li, C.; Liu, J.; Tao, J.; Liu, X.; Zhao, D.; Yin, W.; Chen, H.; Cheng, C.; et al. 2021. Accurate diagnosis of pulmonary nodules using a noninvasive DNA methylation test. *The Journal of Clinical Investigation*, 131(10).

Liang, W.; Zhao, Y.; Huang, W.; Gao, Y.; Xu, W.; Tao, J.; Yang, M.; Li, L.; Ping, W.; Shen, H.; et al. 2019. Non-

invasive diagnosis of early-stage lung cancer using high-throughput targeted DNA methylation sequencing of circulating tumor DNA (ctDNA). *Theranostics*, 9(7): 2056.

Macías-García, L.; Martínez-Ballesteros, M.; Luna-Romera, J. M.; García-Heredia, J. M.; García-Gutiérrez, J.; and Riquelme-Santos, J. C. 2020. Autoencoded DNA methylation data to predict breast cancer recurrence: Machine learning models and gene-weight significance. *Artificial Intelligence in Medicine*, 110: 101976.

Ni, P.; Huang, N.; Zhang, Z.; Wang, D.-P.; Liang, F.; Miao, Y.; Xiao, C.-L.; Luo, F.; and Wang, J. 2019. DeepSignal: detecting DNA methylation state from Nanopore sequencing reads using deep-learning. *Bioinformatics*, 35(22): 4586–4595.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32*, 8024–8035. Curran Associates, Inc.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830.

Sandoval, J.; Heyn, H.; Moran, S.; Serra-Musach, J.; Pujana, M. A.; Bibikova, M.; and Esteller, M. 2011. Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics*, 6(6): 692–702.

Schaffer, J. 2015. What not to multiply without necessity. *Australasian Journal of Philosophy*, 93(4): 644–664.

Shoemaker, R.; Deng, J.; Wang, W.; and Zhang, K. 2010. Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. *Genome research*, 20(7): 883–889.

Sun, L.; Namboodiri, S.; Chen, E.; and Sun, S. 2019. Preliminary Analysis of Within-Sample Co-methylation Patterns in Normal and Cancerous Breast Samples. *Cancer informatics*, 18: 1176935119880516.

Tian, Q.; Zou, J.; Tang, J.; Fang, Y.; Yu, Z.; and Fan, S. 2019. MRCNN: a deep learning model for regression of genome-wide DNA methylation. *BMC genomics*, 20(2): 1–10.

Titus, A. J.; Bobak, C. A.; and Christensen, B. C. 2018. A new dimension of breast cancer epigenetics. In *9th International Conference on Bioinformatics Models, Methods and Algorithms*.

Torre, L. A.; Siegel, R. L.; and Jemal, A. 2016. Lung cancer statistics. *Lung cancer and personalized medicine*, 1–19.

Wang, F.; Xu, H.; Zhao, H.; Gelernter, J.; and Zhang, H. 2016. DNA co-methylation modules in postmortem prefrontal cortex tissues of European Australians with alcohol use disorders. *Scientific reports*, 6(1): 1–11.

Wang, Z.; and Wang, Y. 2018. Exploring dna methylation data of lung cancer samples with variational autoencoders. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 1286–1289. IEEE.

Xie, H.; Wang, M.; De Andrade, A.; Bonaldo, M. d. F.; Galat, V.; Arndt, K.; Rajaram, V.; Goldman, S.; Tomita, T.; and Soares, M. B. 2011. Genome-wide quantitative assessment of variation in DNA methylation patterns. *Nucleic acids research*, 39(10): 4099–4108.