A deep embedded refined clustering approach for breast cancer distinction based on DNA methylation

Rocío del Amor[1] · Adrián Colomer[1] · Carlos Monteagudo[2] · Valery Naranjo[1]

Abstract

Epigenetic alterations have an important role in the development of several types of cancer. Epigenetic studies generate a large amount of data, which makes it essential to develop novel models capable of dealing with large-scale data. In this work, we propose a deep embedded refined clustering method for breast cancer differentiation based on DNA methylation. In concrete, the deep learning system presented here uses the levels of CpG island methylation between 0 and 1. The proposed approach is composed of two main stages. The first stage consists in the dimensionality reduction of the methylation data based on an autoencoder. The second stage is a clustering algorithm based on the soft assignment of the latent space provided by the autoencoder. The whole method is optimized through a weighted loss function composed of two terms: reconstruction and classification terms. To the best of the authors' knowledge, no previous studies have focused on the dimensionality reduction algorithms linked to classification trained end-to-end for DNA methylation analysis. The proposed method achieves an unsupervised clustering accuracy of 0.9927 and an error rate (%) of 0.73 on 137 breast tissue samples. After a second test of the deep-learning-based method using a different methylation database, an accuracy of 0.9343 and an error rate (%) of 6.57 on 45 breast tissue samples are obtained. Based on these results, the proposed algorithm outperforms other state-of-the-art methods evaluated under the same conditions for breast cancer classification based on DNA methylation data.

Keywords Deep embedded refined clustering · Breast cancer · DNA methylation · Dimensionality reduction

1 Introduction

Epigenetic mechanisms are crucial for the normal development and maintenance of tissue-specific gene expression profiles in mammals. Recent advances in the field of cancer epigenetics have shown extensive reprogramming of every component of the epigenetic machinery including DNA methylation, histone modifications, nucleosome positioning and noncoding RNAs [24]. In concrete, several studies demonstrate that DNA methylation (DNAm) plays a crucial role in the tumorigenesis process [1, 34].

In mammalian cells, DNA methylation is based on the selective addition of a methyl group to the cytosine nucleotide under the action of DNA methyltransferases [30], as shown in Fig. 1. Specifically, DNAm takes place in cytosines that precede guanines, known as CpG dinucleotides [8].

CpG sites are not randomly distributed throughout the genome, but there are CpG-rich areas known as CpG islands often located in the gene promoting regions. CpG islands are usually largely unmethylated in normal cells. The methylation of these CpG sites silences the promoter activity and correlates negatively with the gene expression. The methylation of the promoter regions in some vital genes, such as tumor suppressor genes, and therefore their inactivation, has been firmly established as one of the most common mechanisms for cancer development [6, 19]. Because the methylation patterns can be observed in the early stages of cancer [24], DNA methylation analysis

✉ Rocío del Amor
  madeam2@upvnet.upv.es

1   Instituto de Investigación e Innovación en Bioingeniería, I3B, Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Spain

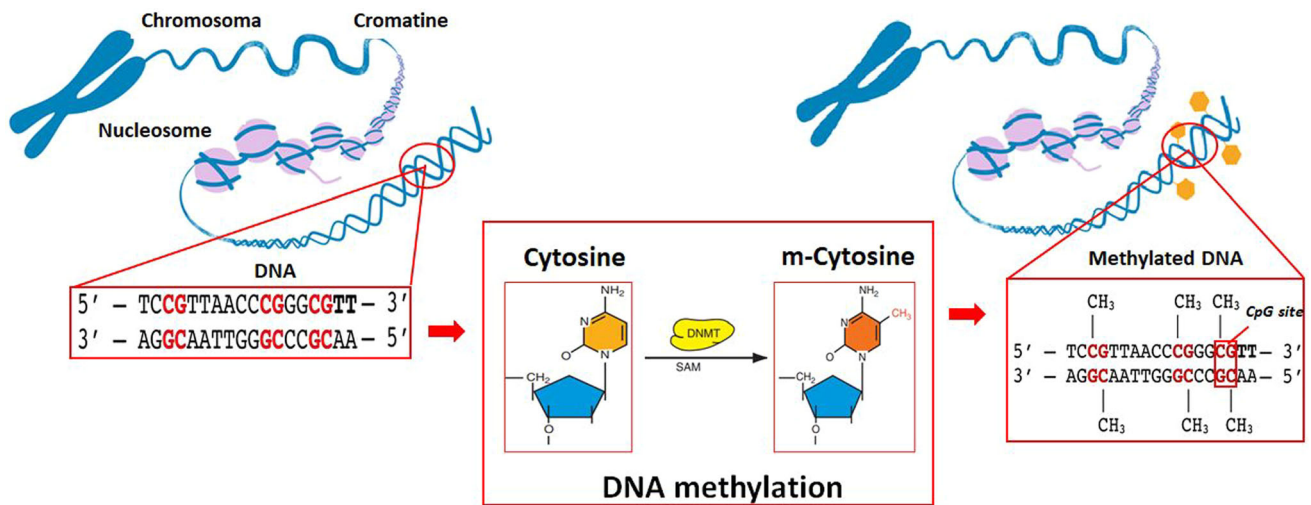2   Pathology Department, Hospital Clínico Universitario de Valencia, Universidad de Valencia, Valencia, Spain

Fig. 1 DNA methylation process. Methylation at 5' position of the cytosine catalyzed by DNMT (DNA methyltransferases) in the presence of S-adenosyl methionine (SAM)

becomes a powerful tool in the early diagnosis, treatment and prognosis of cancer.

The DNA methylation analysis has experienced a revolution during the last decade, especially due to the adaptation of microarray technology to the study of methylation and the emergence of next-generation sequencing (NGS) [4, 21]. These technological advances combined with the development of techniques such as reduced representation bisulfite sequencing (RRBS), which is an efficient and high-throughput approach for analyzing the genome-wide methylation profiles, have allowed the DNA methylation analysis at the molecular level [18]. This is the reason why current methylation studies generate a large amount of data. Additionally, since the study of DNA methylation is still a bit expensive, the number of available samples is relatively low. The extremely high dimensions of the methylation data compared to the generally small number of available samples leads to the so-called curse of dimensionality problem, the main limitation in the development of appropriate methods for DNAm data analysis.

The curse of dimensionality (COD) was introduced by Belman in 1957 [3] and refers to the difficulty of finding hidden structures when the number of variables is large. The high data dimensionality has different adverse effects: increased computational effort, large waste of space, overfitting and poor visualization [31]. In most cases, a dimensional increase has no significant benefit, since a lower data dimensionality might contain more relevant information. In machine learning problems, a small increase in data dimensionality requires a large increase in data volume to maintain a similar level of performance on tasks such as clustering and regression. A well-established

settlement to mitigate the curse of dimensionality is to transform the data from a higher-dimensional space to a more useful lower-dimensional space [31]. There are different types of dimensionality reduction algorithms. Some of them, such as principal component analysis (PCA) or manifold learning, use linear or nonlinear combinations of existing features to create new features. Others, such as forward selection or random forests, only keep the most important features in the dataset and remove redundant ones.

More recently, several state-of-the-art approaches based on high-dimensional data clustering have been proposed to deal with the curse of dimensionality problem. Among these approaches, stand out the methods based on divisive hierarchical clustering [14, 27] and subspace clustering algorithms [2, 5]. Tasoulis et al. introduced a new approach to divisive hierarchical clustering identifying clusters in nonlinear manifolds. This approach uses the isometric mapping (Isomap) to recursively embed subsets of data in one dimension and then perform a binary partition designed to avoid the splitting of clusters [27]. In [14], the authors proposed a new divisive hierarchical clustering method, in which each partition in the hierarchy is induced by a hyperplane separator. Subspace clustering splits the data samples into groups such that each group contains only data samples lying in the same low-dimensional subspace of the given high-dimensional feature space. In [5], the author proposed a local extension of the well-known iterative subspace clustering algorithms in which the entire cluster is approximated with a single linear/affine subspace. Araújo et al. introduced a soft-subspace clustering algorithm, a self-organizing map (SOM) with a time-

varying structure, to cluster data without any prior knowledge of the number of categories or of the neural network topology, both determined during the training process [2].

To mitigate the curse of dimensionality problem existing in the DNAm data, a dimensionality reduction is necessary before implementing any algorithm that identifies the presence of cancer using methylation profiles [33]. In this context, Yuvaraj et al. presented different algorithms for dimensionality reduction based on PCA and Fisher criterion [33]. However, the DNA methylation datasets cannot be efficiently described by these dimensionality reduction methods due to its non-Gaussian character. Jazayer et al. used a nonnegative matrix factorization (NMF) for the dimensionality reduction of breast methylation data following by ELM and SVM classifiers for cancer identification. However, with the NMF algorithm, it is not possible to directly transfer the input to a smaller dimensional space than the number of samples because this method transfers the data to an output space with a dimension equal to the minimum of {samples, DNAm dimension}. That is the reason why, in this study, the authors use a column-splitting method to overcome the curse of the dimensionality problem [16].

Recent advances in the field of artificial intelligence have allowed the development of deep learning algorithms that perform an embedding of CpG methylation states to extract biologically significant lower-dimensional features [17, 25, 29]. Zhongwei et al. presented a stack of random Boltzmann machine (RBM) layers with the aim of reducing the dimensionality of a breast DNAm set composed of cancer and non-cancer samples. The proposed model first selected the best 5000 features based on variance from over 27,000 features and subsequently used four RBM layers to reduce the number of features to 30. After reducing the data dimensionality, they carried out a binary classification of the generated features using unsupervised methods [25]. Khwaja et al. proposed a deep autoencoder system for differentiation of several cancer types (breast cancer, lung carcinoma, lymphoblastic leukemia and urological tumors) based on the DNA methylation states. After a statistical analysis, in which the features providing non-useful information for differentiation between cancer classes are eliminated, the authors used a deep belief network for dimensionality reduction with a posterior supervised classification [17]. Titus et al. proposed an unsupervised deep learning framework with variational autoencoders (VAEs) to learn latent representations of the DNA methylation from three independent breast tumor datasets. They demonstrate the feasibility of VAEs to track representative differential methylation patterns among clinical subtypes of breast tumors, but they do not perform any classification with the extracted characteristics [29].

Several state-of-the-art methods propose different unsupervised and supervised classification algorithms for cancer identification after performing a dimensionality reduction of the DNA methylation data [16, 17, 19, 25]. However, to the best of the author's knowledge, no previous studies have been focused on the development of both tasks simultaneously. Novel deep learning algorithms have emerged optimizing the dimensionality reduction with unsupervised classification at the same time. These methods, called deep clustering algorithms, have outperformed the state-of-the-art results for different tasks as image classification [11, 12, 32], image segmentation [7], speech separation [13, 23] or RNA sequencing [28]. Therefore, our hypothesis is that since these algorithms perform well with high-dimensional data, they are likely to perform well for methylation data.

For all of the above, in this work, we proposed a deep embedded refined clustering to distinguish cancer thought DNA methylation data. In concrete, this work is developed using two public databases containing DNAm data from breast tissues with and without cancer. The proposed method is composed of an autoencoder to carry out the dimensionality reduction followed by a soft-assignment algorithm to perform an unsupervised classification. This algorithm is end-to-end trained to accomplish the data classification while optimizing the dimensionality reduction. As the main novelty, the method is optimized through a weighted loss function. This loss function is composed of two terms: (1) a reconstruction term in charge of optimizing the latent features provided by the autoencoder algorithm and (2) a clustering term used to improve the classification based on the latent features of the autoencoder. To the best of the authors' knowledge, no previous studies have addressed the distinction of cancer based on DNAm using an end-to-end trained dimensionality reduction and classification method. The proposed method is widely validated and compared to the use of autoencoder and variational autoencoder for dimensionality reduction with a subsequent unsupervised classification.

The rest of the paper is organized as follows: In Sect. 2, we introduce the databases used in this work, DNAm sets containing the methylation level (between 0 and 1) of different CpG region related to cancer. In Sect. 3, we describe the methodology. In particular, Sect. 3.1 describes the statistical analysis performed on the DNAm data, Sect. 3.2 presents the dimensionality reduction algorithms used in this work, conventional and variational autoencoder, and Sect. 3.3 describes the details of the proposed deep clustering method. In Sect. 4, we describe the performed experiments in order to validate our method and in Sect. 5 we discuss the results obtained. Finally, Sect. 6 summarizes the conclusions extracted with the carried out experiments.

2 Materials

For this study, we used two methylation datasets obtained from Gene Expression Omnibus (GEO) website [10]. GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Specifically, we used the GSE32393 and the GSE57285 series to evaluate the proposed method. Note that the methodology proposed in this work was applied on the first series. The second series was used as an external database to perform an additional test and demonstrate the robustness of the proposed methodology. The GSE50220 series is composed of breast tissue samples from 114 breast cancers and 23 non-neoplastic breast tissues. The breast cancer tissue samples come from women (mean age 59.4) who were diagnosed with breast cancer. Among the cancers, 33 were at stage 1 and 81 at stage 2/3/4. All 23 non-neoplastic samples are from healthy women (mean age 47.6). The GSE50220 series is composed of breast tissue samples from 39 breast cancer and 9 normal control. Among the breast cancer, 20 were non-irradiated breast cancer and the rest were irradiated tumors. In all cases, to obtain the methylation data, the Illumina Infinium 27k Human DNA methylation Beadchip v1.2 was used at approximately 27,000 CpGs from women with and without breast cancer. For each sample, 27,578 DNA methylation profiles were obtained. The methylation status of each CpG site varies from 0 to 1. Under ideal conditions, a value of 0 means the CpG site is completely unmethylated and the value of 1 indicates the site is fully methylated.

3 Methodology

3.1 Statistical analysis

To conduct the prescreening procedure and obtain the methylation sites with the most differential methylation expression, a previous statistical analysis of the CpG methylation data was carried out. First, a hypothesis contrast to analyze the level of independence between pairs of variables was performed. For this purpose, the correlation coefficient ρ and the *p-value* of the correlation matrix were calculated to remove those variables that meet both *p-value* $\leq \alpha$ and $|\rho| \leq 0.90$, being α the level of significance with a value of 0.05 for this application. After that, we performed different contrasting hypotheses to analyze the discriminatory ability of each variable regarding the class. Depending on if the variables fit a normal distribution or not, the hypothesis test performed was the t-student or the Wilcoxon rank sum, respectively. After the statistical analysis, we reduced the 27,578 DNA methylation features

of the GSE32393 series to 10,153. These features were the input for the following stage.

3.2 Dimensionality reduction

In order to explore the well-known non-supervised algorithms to reduce the data dimensionality based on deep learning techniques, the conventional and the variational autoencoder were tested. In this section, we detail the characteristics of both algorithms and their main differences.

3.2.1 Conventional autoencoder

Autoencoder (AE) is one of the most significant algorithms in unsupervised data representation. The objective of this method is to train a mapping function to ensure the minimum reconstruction error between input and output [22]. As it can be observed in Fig. 2, the conventional autoencoder architecture is composed mainly of two stages: the encoder and the decoder stages. The encoder step is in charge of transforming the input data $\mathbf{X}$ into a latent representation $\mathbf{Z}$ through a nonlinear mapping function, $\mathbf{Z} = f_\phi(\mathbf{X})$, where ϕ are the learnable parameters of the encoder architecture. The dimensionality of the latent space $\mathbf{Z}$ is much smaller than the corresponding input data to avoid the curse of dimensionality [32]. Since the latent space is a nonlinear combination of the input data with smaller dimensionality, it can represent the most salient features of the data. The decoder stage produces the reconstruction of the data based on the features embedded in the latent space, $\mathbf{R} = g_\theta(\mathbf{Z})$. The reconstructed representation $\mathbf{R}$ is required to be as similar to $\mathbf{X}$ as possible. Therefore, given a set of data samples $\mathbf{X} = \{x_i, ..., x_n\}$, being n the number of available samples, the autoencoder model is optimized with the following formula:
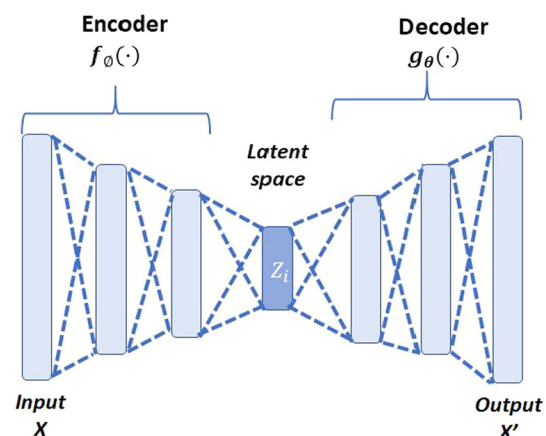

Fig. 2 Architecture of the proposed conventional autoencoder used for the non-supervised dimensionality reduction

$$\min_{\theta,\phi} L_{rec} = \min \frac{1}{n} \sum_{i=1}^{n} ||x_i - g_\theta(f_\phi(x_i))||^2 \qquad (1)$$

where θ and ϕ denote the parameters of encoder and decoder, respectively.

The autoencoder architecture can vary between a simple multilayer perceptron (MLP), a long short-term memory (LSTM) network or a convolutional neural network (CNN), depending on the use case. In case the input data are 1-D and unrelated in time, both the encoder and decoder are usually constructed by a multilayer perceptron.

3.2.2 Variational autoencoder

Variational autoencoder (VAE) is an unsupervised approach composed also of an encoder-decoder architecture like the conventional autoencoder aforementioned [29]. However, the main difference between a conventional and a variational autoencoder lies in the fact that the VAE introduces a regularization into the latent space to improve its properties. With a VAE, the input data are coded as a normal multivariate distribution $p(z|x)$ around a point in the latent space. In this way, the encoder part is optimized to obtain the mean and covariance matrix of a normal multivariate distribution: see Fig. 3.

The VAE algorithm assumes that there is no correlation between any latent space dimensions and, therefore, the covariance matrix is diagonal. In this way, the encoder only needs to assign each input sample to a mean and a variance vectors. In addition, the logarithm of the variance is assigned, as this can take any real number in the range $(-\infty, \infty)$, matching the natural output range from a neural network, whereas that variance values are always positive; see Fig. 4.

In order to provide continuity and completeness to the latent space, it is necessary to regularize both the logarithm of the variance and the mean of the distributions returned by the encoder. This regularization is achieved by matching the encoder output distribution to the standard normal distribution ($\mu = 0$ and $\sigma = 1$).

After obtaining and optimizing the parameters of mean and variance of the latent distributions, it is necessary to take samples of the learned representations to reconstruct the original input data. Samples of the encoder output distribution are obtained as follows:

$$Z \approx p(z|x) = \mu + \sigma \cdot \epsilon \qquad (2)$$

where ϵ is randomly sampled from a standard normal distribution and $\sigma = \exp\left(\frac{\log(\sigma^2)}{2}\right)$.

The minimized loss function in a variational autoencoder is composed of two terms: (1) a reconstruction term that compares the reconstructed data to the original input in order to get as effective encoding–decoding as possible and (2) a regularization term in charge of regularizing the latent space organization, as shown in Fig. 4. The regularization term is expressed as the *Kullback–Leibler* (KL) divergence that measures the difference between the predicted latent probability distribution of the data and the standard normal distribution in terms of mean and variance of the two distributions [9]:

$$D_{KL}[N(\mu, \sigma)||N(0, 1)] = \frac{1}{2} \sum (1 + log(\sigma^2) - \mu^2 - \sigma^2) \qquad (3)$$

The *Kullback–Leibler* function is minimized to 0 if $\mu = 0$ and $log(\sigma^2) = 0$ for all dimensions. As these two terms begin to differ from 0, the variational autoencoder loss increases. The compensation between the reconstruction

Fig. 3 Main differences between a conventional and a variational autoencoder. Instead of just learning a function representing the data (a compressed representation) like conventional autoencoders, variational autoencoders learn the parameters of a probability distribution representing the input data
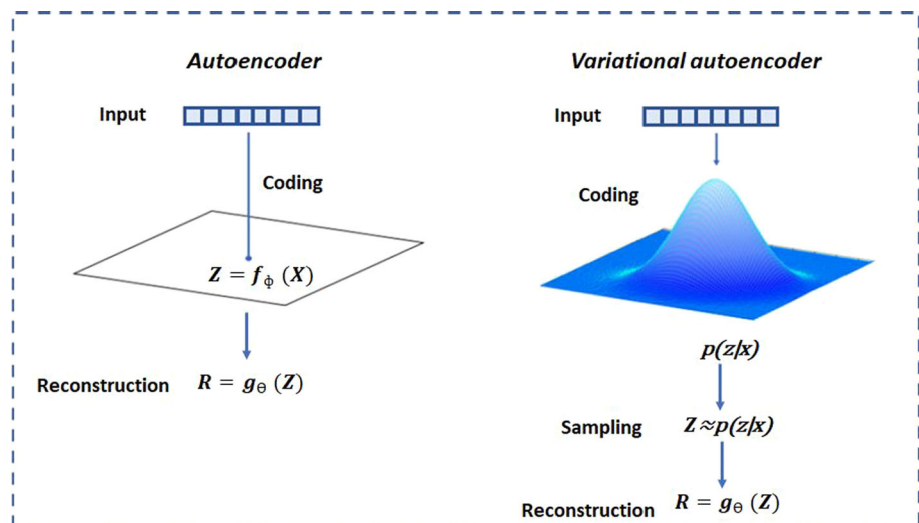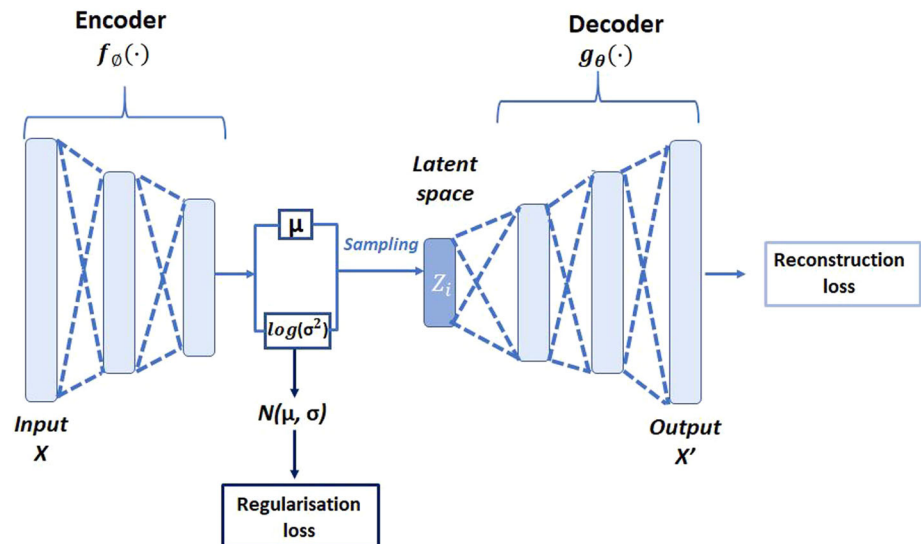
Fig. 4 Architecture of a variational autoencoder. The proposed algorithm is optimized by minimizing two loss functions. One of them corresponding to the latent space regularization and the other one corresponding to the input data reconstruction

error and the KL divergence is a hyperparameter to be adjusted in this type of architecture.

3.3 Proposed method: deep embedded refined clustering

Once the data dimensionality is reduced, we classify the samples in cancerous and non-cancerous. Reducing the data dimensionality without information about the different subjacent data distributions weakens the representativeness of the embedded features concerning the class. and thereby, the performance of the subsequent classification worsens. For this reason, we consider that dimensionality reduction and classification should be optimized at the same time. In this context, we propose a deep embedded refined clustering (DERC) approach for classifying the DNA methylation data (see Fig. 5). It is composed of an autoencoder in charge of the dimensionality reduction and a cluster assignment corresponding to the unsupervised classification stage (clustering layer in Fig. 5). This approach is trained end-to-end optimizing the dimensionality reduction and the unsupervised classification in the same step and not in two different steps as all the algorithms proposed for DNA methylation analysis in the literature.

During the training process, the encoder and decoder weights of the autoencoder, W and $W\prime$, respectively, are updated in each iteration in order to refine the latent features of the encoder output $\mathbf{Z}$. The proposed clustering layer (linked to the encoder output) obtains the soft-assignment probabilities $q_{i,j}$ between the embedded points z_i and the cluster centroids $\{\mu_j\}_{j=1}^k$ every T iterations, being k the number of cluster centroids. The soft-assignment probabilities ($q_{i,j}$) are obtained with the Student's t-distribution proposed in [32]. Using $q_{i,j}$, the target probabilities

$p_{i,j}$ are updated (see Algorithm 1). These target probabilities allow the refinement of the cluster centroids by learning from the current high-confidence assignments. To take into account the refinement of the latent space carried out by autoencoder while the samples are classified in one of the two clusters (cancer and non-cancer), the proposed model is trained end-to-end minimizing both reconstruction L_{rec} and clustering loss L_{cluster} terms:

$$L = L_{\mathrm{cluster}} + \beta L_{\mathrm{rec}} \tag{4}$$

where β balances the importance of the losses due to the reconstruction of the data. The term L_{rec} was defined in Eq. (1), and it is minimized to obtain the maximum similarity between the input and the output data improving the representation of the latent space. L_{cluster} is defined by the Kullback–Leibler (KL) divergence loss between the soft-assignments and the target probabilities, $q_{i,j}$ and $p_{i,j}$, respectively:

$$L_{\mathrm{cluster}} = \sum_i \sum_j p_{i,j} \log \frac{p_{i,j}}{q_{i,j}} \tag{5}$$

The clustering term is minimized to achieve the soft-assignments $q_{i,j}$ and the target $p_{i,j}$ probabilities to be as similar as possible. In this way, the centroids are refined and the latent space obtained by the autoencoder is regularized to achieve a correct distinction between breast cancer and non-breast cancer samples. As discussed above, the hyperparameter β balances the importance of losses due to the data reconstruction. If β is high, the data reconstruction term will predominate and the classification between cancerous and non-cancerous samples will worsen. Otherwise, if this term is too low, the reconstruction losses will be marginal and the features of the latent space will not be optimized correctly. Consequently, the latent

features will be very different from the input data, decreasing the accuracy of the unsupervised classification. Therefore, β is a hyperparameter that needs to be properly adjusted. In Step 2 of Algorithm 1, the methodology used to optimize the proposed DERC algorithm is detailed.

Note that to train the proposed method, a previous initialization of the centroids with latent characteristics is necessary (Step 1 of the Algorithm 1). In the experimental section, we present an experiment (Sect. 4.1.1) aimed at determining which of the dimensionality reduction models is optimal for this initialization.

Algorithm 1: Proposed methodology for the DERC approach.

Input: Methylation data $\mathbf{X}$; number of clusters k; update interval T; batch-size bs; learning rate lr; number of samples n.

Output: Cluster assignment $\{c_i\}_{i=1}^n$ of each methylation sample $\{x_i\}_{i=1}^n$.

Step 1: Previous data dimensionality reduction
(1) Pre-train the proposed autoencoder algorithm.
(2) Obtain the centroid initialisation $\{\mu_j\}_{j=1}^k$ by running K-means on the latent space $\mathbf{Z}$ of the pre-trained autoencoder.

Step 2: Clustering using the proposed DERC method
End-to-end DERC optimization:
for $ite \leftarrow 1$ **to MAXiter do**
  %Choose a batch of samples $X_{bs} \subset \mathbf{X}$
  if $ite\%T == 0$ **then**
    $z_i \leftarrow f_\phi(x_i), \forall x_i \in \mathbf{X}$;
    update $q_{i,j} \leftarrow \frac{(1+||z_i-\mu_j||^2)^{-1}}{\sum_{j'}(1+||z_i-\mu_j||^2)}, j \neq j'$;
    update $p_{i,j} \leftarrow \frac{q_{i,j}^2/f_j}{\sum_{j'} q_{i,j'}^2/f_{j'}}, f_j = \sum_i q_{i,j}$;

  % Update encoder weights W, decoder weights W' and centroids $\{\mu_j\}_{j=1}^k$:
  update $W \leftarrow W - \frac{lr}{bs}\sum_{i=1}^{bs}\left[\beta \frac{\partial L_{rec}}{\partial W'} + \frac{\partial L_{cluster}}{\partial W}\right]$;
  update $W' \leftarrow W' - \frac{lr}{bs}\sum_{i=1}^{bs} \frac{\partial L_{rec}}{\partial W'}$;
  update $\mu_j \leftarrow \mu_j - \frac{lr}{bs}\sum_{i=1}^{bs} \frac{\partial L_{cluster}}{\partial \mu_j}$;

Final prediction stage:
for $i \leftarrow 1$ **to n do**
  $c_i = argmax_j(q_{i,j})$

---

4 Experimental results

As we mentioned in Sect. 2, the DNA methylation databases used were obtained from the Gene Expression Omnibus (GEO) website. In this section, we used the dataset GSE32393 to evaluate the dimensionality reduction and the unsupervised deep clustering performance. The dataset GSE50220 was used as an external validation to demonstrate that the proposed method can generalize to other breast methylation databases. It should be noticed that all experiments were performed on an Intel i7 @ 3.10 GHz of 16 GB of RAM with a Titan V GPU of 12 GB of

RAM. The proposed methods were executed in Python 3.5 using TensorFlow 2.0.

## 4.1 GSE32393 series: performance evaluation

4.1.1 Dimensionality reduction and unsupervised classification separately

As mentioned above, an initial latent space with a lower dimensionality than the input data for the cluster centroid initialization is necessary. In this section, we detail a comparison between the latent space obtained using the conventional and the variational autoencoder and the unsupervised classification results after applying the K-means algorithm on each latent space. In this way, it will be demonstrated which algorithm is the most suitable for dimensionality reduction in the end-to-end proposed method.

Ablation experiment. The 10,153 CpG sites obtained after statistical analysis of the raw methylation data were the input of the proposed dimensionality reduction algorithms, conventional and variational autoencoders. In both cases, the dimensionality reduction was carried out using an architecture composed of 4 stacks. The number of neurons (input, output) of the 3 top layers was set to $\{(10153, 2000), (2000, 500), (500, 70)\}$, respectively (see Fig. 6). These layers were composed of a dense layer with ReLU as activation function except for the last decoder layer that was constituted of the sigmoid function in order to obtain an output value between 0 and 1, range of the methylation data values. The kernel weights were initialized with random numbers drawn from a uniform distribution within $[-l, l]$, where $l = \sqrt{3 \cdot s/n_{\text{input}}}$, being $s = \frac{1}{3}$ and n_{input} the number of input units. The top layer output (latent space dimension) was set to $\{10, 20, 30\}$. Note that, these settings were obtained from empirical evaluations with a wide range of settings and we use only the best parameters here. After intense experiments, the optimal dimension of the latent space for both algorithms turned out to be 10 neurons.

To show the performance of AE and VAE and to demonstrate that they are not over-adjusted to the data, a first experiment was performed using 10% of the GSE32393 database as a validation set and the rest 90% as a training set. Subsequently, both algorithms were trained using the whole database (Entire prediction). The optimal hyperparameters combination was achieved by training both algorithms during 300 epochs, using the stochastic gradient descent (SGD) optimizer with a learning rate of 1 and a batch-size of 8. Regarding the loss function, in the case of the conventional autoencoder, the mean square error (MSE) was used. However, the variational

Fig. 5 Architecture of the proposed method (DERC) to detect breast cancer using DNA methylation data. The proposed algorithm is trained minimizing both, clustering and reconstruction loss
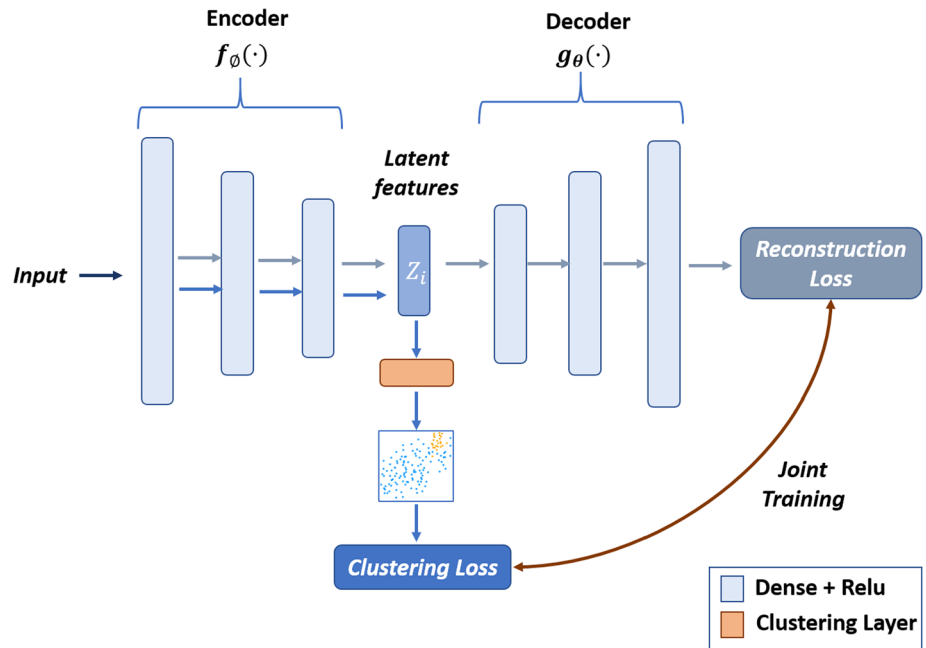
autoencoder loss function was composed of two terms: MSE weighted by 0.8 and Kullback–Leibler (KL) divergence.

After training the dimensional reduction algorithms with the entire GSE32393 series and obtaining the features in the embedding space (encoder output), the classification results were obtained using K-means. To achieve this unsupervised classification, we ran K-means with 80 restarts and selected the best solution.

Qualitative and quantitative results. After training the proposed dimensionality reduction algorithms, the results in terms of reconstruction error for both autoencoders are shown in Table 1.

In order to visualize in a qualitative way the effect of the tested dimensionality reduction methods (AE and VAE) over the data distribution, we used the t-distributed stochastic neighbor embedding (t-SNE) method to represent the latent space into a two-dimensional space. T-SNE is a nonlinear dimensionality reduction technique that embeds high-dimensional data into a space of two or three dimensions, which can then be visualized by a scatter plot [20]. In Fig. 7, we show the representation of the data (latent space of the pre-trained variational autoencoder (a) and latent space of the pre-trained conventional autoencoder (b)) in a two-dimensional space.

To quantitatively evaluate the performance of the clustering assignments, several metrics were computed: the unsupervised clustering accuracy (ACC) [22], the error rate (ER), the false positive (FP) and the false negative (FN) ratios, the adjusted rand index (ARI) [15] and the normalized mutual information (NMI) [26]. The ACC metric is defined as follows:

$$ACC = \max_m \left(\frac{\sum_{i=1}^n 1\{y_i = m(c_i)\}}{n} \right) \tag{6}$$

where y_i is the ground-truth label, c_i is the cluster assignment generated by the algorithm, m is a mapping function which ranges over all possible one-to-one mappings between assignments and labels and n is the total number of samples. The error rate (%) is calculated according to the following formula:

$$ER\ (\%) = (1 - ACC) \cdot 100 \tag{7}$$

The adjusted rand index is defined as follows:

$$ARI = \frac{RI - E[RI]}{\max(RI) - E[RI]} \tag{8}$$

where $RI = \frac{TP+TN}{TP+FN+TN+FP}$, TP and TN are true positive and true negative ratios and E[RI] is the expected index. The ARI can yield negative values if index (RI) is less than expected index E[RI].

The normalized mutual information is defined by the following formula:

$$NMI = \frac{\sum_{i=1}^k \sum_{j=1}^k n_{i,j} \log \frac{n_{i,j}}{n_i \cdot \hat{n}_j}}{\sqrt{\left(\sum_{i=1}^k n_i \log \frac{n_i}{n}\right)\left(\sum_{j=1}^k \hat{n}_j \log \frac{\hat{n}_j}{n}\right)}} \tag{9}$$

where $n_{i,j}$ denotes the number of data points which are in the intersection between cluster c_i and class y_i, n_i is the number of data points in cluster c_i and $\hat{n}_j$ is the number of data points in class y_j.

In Table 2, the abovementioned metrics were calculated for the input data + K-means clustering, the latent space of the pre-trained autoencoder (AE) + K-means and the
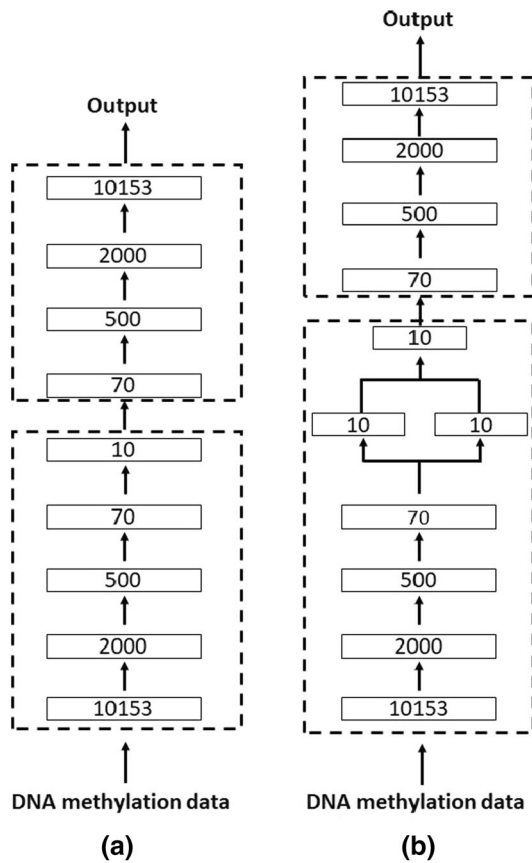
Fig. 6 Final dimensionality reduction architectures. **a** Conventional autoencoder architecture composed of 4 stacks. **b** Variational autoencoder architecture composed of 4 stacks. Note that with the variational autoencoder algorithm, the latent space is obtained in two stages, two dense layers of 10 neurons representing the mean and the logarithm of the variance of the latent distribution and a sampling layer to obtain the points of the latent space

Table 1 Reconstruction error of the proposed dimensionality reduction algorithms

Method	Reconstruction loss		
	Training	Validation	Entire prediction
AE	**0.0062**	**0.0054**	**0.0057**
VAE	0.0082	0.0074	0.0082

Conventional autoencoder (AE) and variational autoencoder (VAE)

The bold values point out the method that has obtained the best results

4.1.1), we added the clustering layer to the output of the autoencoder latent space (see Table 3) for the layer layout in the final architecture.

In order to evaluate the β value on the performance of the clustering algorithm, we kept the rest of the hyperparameters constant during the different experiments. In particular, the entire deep embedding clustering method was optimized by stochastic gradient descent (SGD) with a learning rate of 0.01 and a momentum of 0.9. The proposed method was trained during 50 epochs using a batch-size of 8 samples and the target distribution of the clustering layer was updated every 10 iterations. Note that these hyperparameters were obtained from empirical evaluations with a wide range of settings. β was a variable parameter and various experiments were conducted by setting its value with $\{0.95, 0.85, 0.75, 0.65\}$.

Quantitative results. In this case, we show the unsupervised classification results provided by the proposed deep embedded refined clustering (DERC) method depending on the β value (see Table 4).

4.2 GSE50220 series: generalization ability of the DERC algorithm

In this section, we expose the results for the prediction of an external test set (see Table 5). The goal of this section is to demonstrate that the proposed DERC method could be valid to perform the feature extraction and unsupervised classification from methylation data. Therefore, we made use of the GSE50220 series as an external test set to check the behaviour of the proposed methods with new breast cancer samples.

4.3 Comparison with the state of the art

In order to provide the superiority of the proposed method for DNAm analysis, we compared our approach with well-known methods for high-dimensional clustering. In concrete, we use the divisive hierarchical clustering methods based on isometric mapping using the maximum distance between consecutive one-dimensional embeddings and the
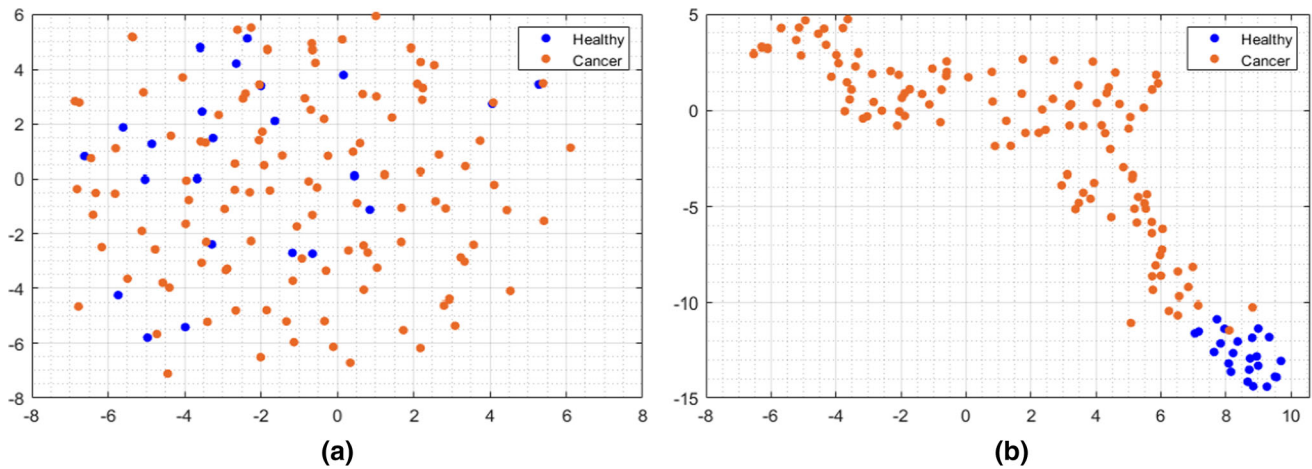
latent space of the pre-trained variational autoencoder (VAE)+ K-means. Note that the input data are referred to the 10,153 features extracted after the statistical analysis.

4.1.2 Dimensionality reduction and unsupervised classification jointly (DERC)

After initializing the centroids using the algorithm with the lowest losses and the better prediction when K-means was used, in this case the conventional autoencoder, the deep embedded refined clustering algorithm was trained. As we explained in Sect. 3.3, the β value, which weights the terms that composed the loss function of the DERC algorithm, is an important parameter to adjust. For this reason, we develop in this section a comparison between different β values exposing their influence on the clustering assignment.

Ablation experiment. After pre-training the conventional autoencoder model (with the parameters detailed in Sect.

Fig. 7 Latent space of the dimensionality reduction algorithms. **a** Visualization of 10-dimensional features extracted by the latent space of the pre-trained variational autoencoder. **b** Visualization of 10-dimensional features extracted by the latent space of the pre-trained conventional autoencoder

Table 2 Comparison of the K-means clustering effect based on different feature extraction

Method	ACC	ER(%)	FN	FP	ARI	NMI
Input Data + K-means	0.6715	32.85	45	0	0.1140	0.2355
AE + K-means	0.9343	6.57	9	0	0.7184	0.6300
VAE + K-means	0.5693	43.07	50	9	0.0133	0.0142
Proposed method	0.9927	0.73	1	0	0.9643	0.9212

Table 3 Architecture of the proposed deep embedded refined clustering model

Layer name	Output shape	Connected to
Input_layer	10153	N/A
Encoder_0	2000	Input_layer
Encoder_1	500	Encoder_0
Encoder_2	70	Encoder_1
Encoder_3	10	Encoder_2
Decoder_3	70	Encoder_3
Decoder_2	500	Decoder_3
Decoder_1	2000	Decoder_2
Clustering_layer	2	Encoder_3
Decoder_0	10153	Decoder_1

Table 4 Comparison of the clustering effect of the proposed DERC based on different β values

Method	ACC	ER(%)	FN	FP	ARI	NMI
DERC ($\beta = 0.95$)	0.9708	2.92	4	0	0.8643	0.7796
DERC ($\beta = 0.85$)	0.9781	2.19	3	0	0.8965	0.8198
DERC ($\beta = 0.75$)	**0.9927**	**0.73**	**1**	**0**	**0.9643**	**0.9212**
DERC ($\beta = 0.65$)	0.9854	1.4600	2	0	0.9298	0.8659

The bold values point out the method that has obtained the best results

5 Discussion

In this work, we present a deep embedded refined clustering approach to automatically detect patients suffering for cancer using DNA methylation data. In concrete, the proposed algorithm was evaluated using two breast methylation datasets.

As it can be observed in Table 2, an optimal data dimensionality reduction is essential to improve the classification results when working with high-dimensional data. Using the K-means algorithm as non-supervised classifier, the dimensionality reduction carried out by the pre-trained AE on the input data (DNAm profiles obtained after statistical analysis) improves the ACC results from 0.6715 to 0.9343. However, with the VAE algorithm, the ACC results do not improve, obtaining a value of 0.5693.

global minimum of the corresponding density estimator (i-DivClu-M and i-DivClu-D, respectively) [27] and the subspace methods based on local affine and convex hull [5] (LSC-aff. hull and LSC-conv. hull, respectively); see Table 6.

Table 5 Results obtained over the external dataset

Method	ACC	ER(%)	FN	FP	ARI	NMI
Input Data + K-means	0.6042	39.58	18	0	0.0445	0.2133
AE + K-means	0.8542	14.57	7	0	0.4727	0.4541
DERC ($\beta = 0.75$)	**0.9375**	**6.25**	**3**	**0**	**0.7374**	**0.6554**

Note that in this case, the input data corresponds to the GSE50220 series after selecting the CpG sites extracted with the statistical analysis in the GSE32393 series

The bold values point out the method that has obtained the best results

As it can be seen in Fig. 7, the latent space of the VAE is centered around 0 due to the regularization effect. This fact makes it impossible to distinguish between the different classes. Additionally, the reconstruction losses obtained by the VAE are higher than those reached by the AE (see Table 1). Therefore, it can be concluded that the conventional autoencoder is the most suitable algorithm to reduce the DNAm dimensionality.

Moreover, regarding the comparison between classifying separately and jointly to the dimensionality reduction, ACC results show an improvement from 0.9343 to 0.9927 when the dimensionality reduction and the unsupervised classification are optimized all at once. In Table 4, it can be observed the effect of the β value on the unsupervised classification results. In this way, it can be demonstrated that when the contribution of the reconstruction losses is too low (low β) or too high (high β), the ACC results are worse compared to a more balanced contribution. However, all the accuracy results shown in Table 4, joint optimization, are higher than those obtained when applying K-means in the autoencoder latent space, separate optimization. Therefore, it is proven that when the classification is carried out at the same time as the dimensionality reduction is optimized, the best results are obtained.

This fact can also be demonstrated when the results of the proposed method are compared to those obtained in the literature [16, 25]. As discussed in Sect. 1, in [25], the authors proposed a dimensionality reduction algorithm followed by several unsupervised classification algorithms.

They applied their methods to the same breast cancer database used in this paper (GSE32393 series). Note that they obtained an error rate of 2.94 using a deep neural network (DNN) following a self-organizing feature map (SOM) compared to 0.73 obtained by the proposed method. Furthermore, their algorithm predicted 4 cancer examples as healthy, while our method only misclassifies one cancer sample. Therefore, it confirms that the proposed deep embedded refined clustering algorithm improves the results when it is applied to DNA methylation data. Additionally, the algorithm proposed in [16] used the same breast cancer database (GSE32393 series). In this case, they used a nonnegative matrix factorization (NMF) for dimensionality reduction following by supervised algorithms for classification. Their main limitation is that, according to the authors, the NMF algorithm cannot directly reduce the number of features (27,578) to a lower dimension than the number of samples (137). Therefore, they used a method called column-splitting in which they separated the original data into different matrices. They could not reduce the original data to a single latent space because they had to reduce each data matrix independently. In this way, the overall information of all original features is not taken into account. They used a K-fold for the algorithm validation and obtaining a 100 % of accuracy when they used 900 and 2700 CpG sites. However, both resulting models were overfitted as it is demonstrated when they reduced the number of features to 540 and the accuracy dropped to 97.85 %, lower than achieved with the proposed method.

Table 6 Results reached by the state-of-the-art approaches in comparison with the proposed method when predicting the DNAm databases. Note that $S1$ refers to the primary set (GSE32393 series) and $S2$ to the external database (GSE50220 series)

	ACC		ER (%)		FN		FP		ARI		NMI	
	$S1$	$S2$	$S1$	$S2$	$S1$	$S2$	$S1$	$S2$	$S1$	$S2$	$S1$	$S2$
i-DivClu-D [27]	0.7600	0.6700	24.00	33.00	33	16	0	0	0.2546	0.1000	0.3113	0.2422
i-DivClu-M [27]	0.8832	0.7917	11.68	20.83	11	10	5	0	0.5149	0.3203	0.3391	0.3618
LSC-aff. hull [5]	0.8248	0.5208	17.52	47.92	24	23	0	0	0.3934	-0.0560	0.3921	0.1545
LSC-conv. hull [5]	0.8248	0.5625	17.52	43.75	24	21	0	0	0.3934	-0.0210	0.3921	0.1700
Proposed	**0.9927**	**0.9375**	**0.73**	**6.25**	**1**	**3**	**0**	**0**	**0.9643**	**0.7374**	**0.9212**	**0.6554**

The bold values point out the method that has obtained the best results

Additionally, the authors of [16] claimed that it is important to reduce the number of features to a smaller space than the total number of examples. However, they were only able to reduce the features to 540 (due to NMF restrictions) which is about 5 times the number of examples they used to train their models.

The results obtained by our proposed method on the external database (GSE50220 series) reported closely similar values to those reached in the primary set (GSE32393 series). This fact indicates that the proposed deep clustering model is perfectly applicable to other breast tissue databases (see Table 5).

Furthermore, to objectively contrast the proposed method with other state-of-the-art high-dimensional clustering approaches, we replicated the experiments performed by [5, 27] with the two DNAm databases proposed in this paper (see Table 6). The results obtained show a clear outperformance of the proposed method with respect to the rest of the state-of-the-art models for all metrics. The methods based on convex hull clustering do not achieve satisfactory results on the DNAm databases due to their strong dependence on initialization. Both methodologies, isometric mapping for divisive clustering (i-Div) and local subspace clustering (LSC), show a decrease in the performance when tested on the external database (S2), demonstrating that they are not scalable for the classification of methylation data, especially when the number of samples is limited.

6 Conclusion

In this paper, a deep embedded refined clustering based on breast cancer classification using DNA methylation data has been presented. To the best of the authors' knowledge, no previous studies using DNA methylation are based on algorithms that can optimize the dimensionality reduction and the classification of the data at the same time. As demonstrated throughout the manuscript, the method proposed in this paper improves the results of algorithms using dimensionality reduction and subsequent classification.

The proposed method allows the breast cancer classification using a latent space of only 10 features, which means a reduction in the dimensionality of 99.9637%. The technology used in this study for data acquisition is the Illumina Infinium 27k Human DNA methylation Beadchip v1.2 which uses probes on the 27k array target regions of the human genome to measure methylation levels at 27,578 CpG dinucleotides in 14,495 genes. As verified through this work, many of the CpG sites obtained in the DNA methylation analysis are not relevant in the breast cancer classification. After ensuring model viability with a larger breast cancer database, the CpG sites from which the level

of methylation is obtained could be reduced decreasing the cost and time of methylation analysis. Therefore, this work could contribute to a faster and more effective diagnosis of breast cancer, improving cancer care and advancing the future of breast cancer research technologies.

From a technical perspective, future lines of work will focus on adapting and applying the proposed method to identify and appropriately classify other challenging disorders, such as melanocytic tumors. In this way, the general applicability of the model for the detection of different types of cancer could be demonstrated.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Akhavan-Niaki H, Samadani AA (2013) DNA methylation and cancer development: molecular mechanism. Cell Biochem Biophys 67(2):501–513
2. Araújo AF, Antonino VO, Ponce-Guevara KL (2020) Self-organizing subspace clustering for high-dimensional and multi-view data. Neural Netw 130:253–268
3. Bellman R (1957) Dynamic programming. Princeton Univ, Princeton
4. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, Delano D, Zhang L, Schroth GP, Gunderson KL et al (2011) High density DNA methylation array with single CPG site resolution. Genomics 98(4):288–295
5. Cevikalp H (2019) High-dimensional data clustering by using local affine/convex hulls. Pattern Recognit Lett 128:427–432

6. Du P, Zhang X, Huang CC, Jafari N, Kibbe WA, Hou L, Lin SM (2010) Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis. BMC Bioinf 11(1):587

7. Enguehard J, O'Halloran P, Gholipour A (2019) Semi-supervised learning with deep embedded clustering for image classification and segmentation. IEEE Access 7:11093–11104

8. Esteller M (2008) Epigenetics in cancer. New Engl J Med 358(11):1148–1159

9. Foster D (2019) Generative deep learning: teaching machines to paint, write, compose, and play. O'Reilly Media

10. GEO: Epigenome analysis of breast tissue from women with and without breast cancer. http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=gse32393

11. Guo X, Liu X, Zhu E, Yin J (2017) Deep clustering with convolutional autoencoders. In: International conference on neural information processing, pp. 373–382. Springer

12. Guo X, Zhu E, Liu X, Yin J (2018) Deep embedded clustering with data augmentation. In: Asian conference on machine learning, pp. 550–565

13. Hershey JR, Chen Z, Le Roux J, Watanabe S (2016) Deep clustering: Discriminative embeddings for segmentation and separation. In: 2016 IEEE international conference on acoustics, apeech and signal processing (ICASSP), pp. 31–35. IEEE

14. Hofmeyr DP (2016) Clustering by minimum cut hyperplanes. IEEE Trans Pattern Anal Mach Intell 39(8):1547–1560

15. Hubert L, Arabie P (1985) Comparing partitions. J Classif 2(1):193–218

16. Jazayeri N, Sajedi H (2020) Breast cancer diagnosis based on genomic data and extreme learning machine. SN Appl Sci 2(1):3

17. Khwaja M, Kalofonou M, Toumazou C (2018) A deep autoencoder system for differentiation of cancer types based on DNA methylation state. arXiv preprint arXiv:1810.01243

18. Laird PW (2010) Principles and challenges of genome-wide DNA methylation analysis. Nature Rev Genet 11(3):191–203

19. Liu B, Liu Y, Pan X, Li M, Yang S, Li SC (2019) DNA methylation markers for pan-cancer prediction by deep learning. Genes 10(10):778

20. Maaten LD, Hinton G (2008) Visualizing data using t-sne. J Mach Learn Res 9(Nov):2579–2605

21. Martorell-Marugán J, Tabik S, Benhammou Y, del Val C, Zwir I, Herrera F, Carmona-Sáez P (2019) Deep learning in omics data analysis and precision medicine. In: Computational Biology [Internet]. Codon Publications

22. Min E, Guo X, Liu Q, Zhang G, Cui J, Long J (2018) A survey of clustering with deep learning: from the perspective of network architecture. IEEE Access 6:39501–39514

23. Prasetio B.H, Tamura H, Tanno K (2019) A deep time-delay embedded algorithm for unsupervised stress speech clustering. In: 2019 IEEE international conference on systems, man and cybernetics (SMC), pp. 1193–1198. IEEE

24. Sharma S, Kelly TK, Jones PA (2010) Epigenetics in cancer. Carcinogenesis 31(1):27–36

25. Si Z, Yu H, Ma Z (2016) Learning deep features for DNA methylation data analysis. IEEE Access 4:2732–2737

26. Strehl A, Ghosh J (2002) Cluster ensembles—a knowledge reuse framework for combining multiple partitions. J Mach Learn Res 3(Dec):583–617

27. Tasoulis S, Pavlidis NG, Roos T (2020) Nonlinear dimensionality reduction for clustering. Pattern Recognit 107:107508

28. Tian T, Wan J, Song Q, Wei Z (2019) Clustering single-cell RNA-seq data with a model-based deep learning approach. Nat Mach Intell 1(4):191–198

29. Titus AJ, Wilkins OM, Bobak CA, Christensen BC (2018) Unsupervised deep learning with variational autoencoders applied to breast tumor genome-wide DNA methylation data with biologic feature extraction. bioRxiv p. 433763

30. Tsou JA, Hagen JA, Carpenter CL, Laird-Offringa IA (2002) DNA methylation analysis: a powerful new tool for lung cancer diagnosis. Oncogene 21(35):5450–5461

31. Venkat N (2018) The curse of dimensionality: Inside out

32. Xie J, Girshick R, Farhadi A (2018) Unsupervised deep embedding for clustering analysis. In: International conference on machine learning, pp. 478–487 x

33. Yuvaraj N, Vivekanandan P (2013) An efficient SVM based tumor classification with symmetry non-negative matrix factorization using gene expression data. In: 2013 International conference on information communication and embedded systems (Icices), pp. 761–768. IEEE

34. Zhang W, Spector TD, Deloukas P, Bell JT, Engelhardt BE (2015) Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements. Genome Biol 16(1):14