A DEEP EMBEDDED FRAMEWORK FOR SPITZOID NEOPLASM CLASSIFICATION USING DNA METHYLATION DATA

Rocío del Amor*, Adrián Colomer*, Carlos Monteagudo †, María José Garzón ‡, José Luis García-Giménez §, and Valery Naranjo*

*Instituto de Investigación e Innovación en Bioingeniería, I3B, Universitat Politècnica de València, Valencia, Spain
Email: {madeam2, adcogra, vnaranjo}@i3b.upv.es

† Pathology Department. Hospital Clínico Universitario de Valencia, Universidad de Valencia, Valencia, Spain.

‡ Consorcio Centro de Investigación Biomédica en Red, Instituto de Salud Carlos III, Instituto de Investigación Sanitaria INCLIVA, Valencia, Spain.

§ Bioinformatics and Computational Diagnostics Unit EpiDisease S.L., Valencia, Spain.

***Abstract*—Spitzoid melanocytic tumors (SMT) are a group of neoplasms that represent a formidable diagnostic challenge for dermatopathologists. DNA methylation (DNAm) is a well-defined epigenetic factor that has an important role in the development of these lesions. In this work, we propose different deep-learning-based approaches to address the Spitzoid neoplasms detection from DNAm. We use an autoencoder and a variational autoencoder for dimensionality reduction with a subsequently supervised classification. Additionally, we present a deep embedded refined clustering algorithm able to optimize the latent space at the same time that the non-supervised classification task is performed. This novel approach in DNAm supposes a step forward in the SMT detection as suggest the obtained results (acc = 0.9). Additionally, making use of the resulting model, we present a subspace-prototypical-based approach for the prognostic prediction of uncertain malignant potential samples, which is nowadays the hottest open area in SMT detection.**

***Index Terms*—Dimensionality reduction, deep embedded refined clustering, DNA methylation, Spitzoid neoplasms.**

I. Introduction

Skin cancer represents the most common group of malignant neoplasms among Caucasians [1], [2]. In fact, nowadays, one in three diagnosed cancers is skin cancer [3]. Although two of the most commonly diagnosed skin cancers are basal and squamous cell carcinoma, non-melanocytic lesions developed from non-pigmented skin cells, the most aggressive and dangerous skin cancer is melanoma [4]. Melanoma is an aggressive neoplasm with numerous mechanisms of resistance against therapeutic agents [5]. In conventional melanocytic tumors, a precise pathological distinction between benign (melanocytic nevus) and malignant (melanoma) is possible. However, there are melanocytic neoplasms which cannot be classified properly with the current histopathological algorithms. This difficulty leads to an underdiagnosis such as nevus, an overdiagnosis such as melanoma or a diagnosis of uncertain malignant potential (UMP), which is the real diagnostic challenge for the pathologists since their prognosis is unknown [6]. Most of this heterogeneous group of melanocytic tumors belong to the Spitzoid tumor category [7]. Spitzoid melanocytic neoplasms are rare skin lesions that constitute diagnostic problems for dermatopathologists on a regular basis [6]. The discrepancy between the histopathologic appearance and its clinical evolution is the reason why molecular studies are needed to understand the biology and predict the clinical behavior of the Spitzoid tumors. These molecular techniques include: fluorescence in situ hybridization (FISH) [8], DNA methylation (DNAm) [9] and miRNAs expression [10]. Specifically, DNAm is a frequent event in melanocytic tumors related to their development and metastatic progression. Therefore, it may serve as a diagnostic, prognostic and therapeutic biomarker [11]–[13]. DNAm is based on the addition of a methyl group to the cytosine nucleotide under the DNA methyltransferases (DNMT) action [14], see Fig. 1. Specifically, DNAm takes place in cytosines that precede guanines, known as CpG dinucleotides [15]. The CpG sites are often located in the gene promoting regions and the methylation of these regions has been firmly established as one of the most common mechanisms for gene regulation in cancer.

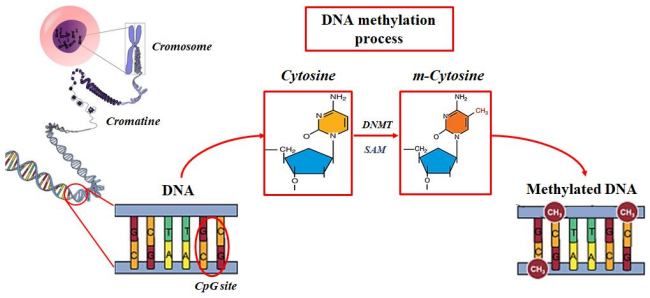

Fig. 1. DNA methylation process. Methylation at position 5' of the cytosine catalyzed by DNMT (DNA methyltransferases) in the presence of S-adenosyl methionine (SAM).

This work has received funding from Horizon 2020, the European Union's Framework Programme for Research and Innovation, under grant agreement No. 860627 (CLARIFY), the Spanish Ministry of Economy and Competitiveness through project PID2019-105142RB-C21 (AI4SKIN) and SICAP (DPI2016-77869-C2-1-R) and GVA through project PROMETEO/2019/109.

The DNA methylation analysis generates a large amount of data compared to the generally small number of samples available. Therefore, a dimensionality reduction is necessary before implementing any automatic classification algorithm. Several studies have proposed different dimensionality reduction and classification algorithms for the cancer identification (e.g breast cancer [16]–[18] and lung cancer [18], [19]) using the DNAm level. In this context, Jazayeri et al. used a non-negative matrix factorization for dimensionality reduction and the extreme learning machine and the support vector machine algorithms for breast cancer classification [16]. Zhongwei et al. presented a stack of Random Boltzmann Machine (RBM) layers to reduce the dimensionality of a breast DNAm set. Subsequently, a binary classification through several unsupervised methods was proposed [17]. Khwaja et al. proposed a deep autoencoder system for differentiation of several cancer types (breast cancer and lung carcinoma) 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 [18]. Zhenxing et al. used a variational autoencoder to reduce the dimensionality of lung DNAm cancer samples. Afterward, a logistic regression classifier on the encoded latent features was proposed [19]. To the best of the author's knowledge, no previous studies have been focused on the Spitzoid neoplasm distinction through the DNAm data.

For all the above, in this paper, we propose and compare two different deep-learning-based approaches to address the SMT distinction between nevus and melanoma using the level of DNAm. Specifically, the first one considers two unsupervised algorithms for dimensionality reduction of DNAm (autoencoder and variational autoencoder) with a subsequently supervised classification. The second one is based on an agnostic learning algorithm that allows reducing the data dimensionality whereas a refinement of the classification is carried out. Furthermore, we propose, for the first time in this kind of data, a subspace prototypical algorithm to determine the prognostic of UMP samples in malignant or nevus Spitzoid neoplasms.

II. Material and methods

A. Material

The experiments detailed in this paper were performed on a private database composed of patient samples collected in the Pathological Anatomy Service of the University Clinical Hospital of Valencia from 1990 to 2017. Specifically, the dataset contains the methylation level, in the range 0-1, of 491 CpG sites from 39 patients with melanocytic Spitzoid lesions. Eight of the thirty-nine patients under study were diagnosed as malignant melanocytic lesions (melanoma), twelve as benign (nevus) and the rest as uncertain malignant potential (UMP). To conduct the prescreening procedure and obtaining the methylation sites with the most differential methylation expression, a statistical analysis of the CpG methylation was carried out before facing the dimensionality reduction stage. 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 491 DNA methylation features of each sample under study to 62. These features were the input for the following stage.

B. CpG dimensionality reduction

In order to overcome the curse of dimensionality problem due to the extremely high dimensions of the DNAm data, a dimensionality reduction was carried. In this study, two unsupervised deep learning frameworks, autoencoder (AE) and variational autoencoder (VAE), were proposed.

- *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 using mainly of two stages: the encoder and the decoder stage [20]. The encoder function is in charge of transforming the input data $\mathbf{X}$ into a latent representation $\mathbf{Z}$ through a non-linear mapping function, $\mathbf{Z} = f_\phi(\mathbf{X})$, where ϕ are learnable parameters of the encoder stage. The dimensionality of the latent space $\mathbf{Z}$ is smaller than this input data to avoid the curse of dimensionality [21]. The decoder stage produces the data reconstruction 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 x_i where $i = 1, ..., n$ being n the number of available samples, the autoencoder model is optimized as follows:

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

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

In this case, the encoder and the decoder stage was composed of three stacks. Specifically, three layers with 62, 20 and 7 neurons, respectively. 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 initialised to random numbers drawn from a uniform distribution within *[-limit,limit]*, where

$limit = \sqrt{3 \cdot scale/n_{input}}$, being $scale = \frac{1}{3}$ and n_{input} the number of input units. Regarding the hyper-parameters combination, Stochastic gradient descent (SGD) optimizer with a learning rate (lr) of 0.001 reported the best learning curves when the model was forward and backpropagated during 900 epochs with a batch size (bs) of 8, minimizing the means square error (MSE) loss function.

- *Variational autoencoder*

Variational autoencoders are based on an encoder and a decoder stage as well as the AE method. The main difference is that VAE models learn the distribution of explanatory features over samples through two latent representations: a mean and standard deviation vector encoding [22]. In this case, the VAE model is optimized by minimizing both Equation (1), to achieve the best input data reconstruction, and a term corresponding to the regularization of the latent space organization. This term is expressed as the *Kulback-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 logarithm:

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

The KL function is minimised 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 error and the KL divergence is a hyper-parameter to be adjusted in this type of architecture. Regarding the architecture and hyperparameters combination used, this was the same as that used for the aforementioned autoencoder. Additionally, the reconstruction loss was weighted by 0.7.

C. Classification of Spitzoid melanocytic neoplasms

1) Supervised methylation classification: After the dimensionality reduction of the DNAm data from patients with malignant and benign melanocytic neoplasms, the classification of these tumors was carried out. In this case, we used a multilayer perceptron (MLP) from the latent space of the proposed reduction algorithms, Fig. 2. The multilayer perceptron architecture consisted of three layers: an input layer of 7 neurons, a hidden layer of 4 neurons in the case of the autoencoder and 3 for the variational autoencoder and an output layer of two neurons with the softmax as the activation function. Table I shows the combination of hyperparameters depending on whether the input of the MLP was the latent space of the autoencoder or the variational autoencoder.

2) Deep Embedded Refined Clustering for methylation classification: In this section, we propose an agnostic learning algorithm for Spitzoid classification. Specifically, we present a deep embedded clustering (DERC) algorithm (see Fig. 3) [23]. It is composed of an autoencoder for dimensionality reduction (with the same architecture as explained in Section II-B)

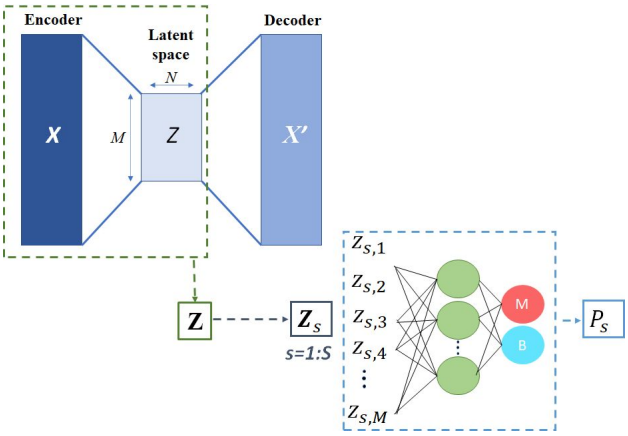

Fig. 2. Workflow for supervised classification of DNAm data, where S is the number of folds, in this case 5.

TABLE I
HYPERPARAMETERS COMBINATION. WBCE: WEIGHTED BINARY CROSS-ENTROPY.

	Optimizer	lr	Epochs	bs	Loss function
AE	SGD	0.01	300	8	WBCE
VAE	SGD	0.1	900	8	WBCE

and a cluster assignment corresponding to the unsupervised classification stage. The clustering process consists of two main stages: (1) a soft assignment between the embedded points (encoder output) and the cluster centroids (initialized using K-means on the encoder output of Section II-B) as follows:

$$q_{i,j} = \frac{(1 + ||z_i - \mu_j||^2)}{\sum_{j'}(1 + ||z_i - \mu_j||^2)} \quad (3)$$

where $q_{i,j}$ is the probability of assigning sample i to cluster j of embedded refined points $z_i = f_\phi(x_i)$ and (2) the refinement of the cluster centroids using the following target distribution:

$$p_{i,j} = \frac{q_{i,j}^2/f_j}{\sum_{j'} q_{i,j'}^2/f_{j'}} \quad (4)$$

with $f_j = \sum_i q_{i,j}$ the soft cluster frequencies and $q_{i,j} \neq q_{i,j'}$.

This approach was trained during 240 epochs with Adadelta as an optimizer using a learning rate of 1 and a batch-size of 8. The target distribution of the clustering layer was updated every 10 iterations. This method optimizes the dimensionality reduction while performing the centroid refinement. Thereby, it uses a loss function L with two terms (see Fig. 3): (1) L_r, MSE for dimensionality reduction (weighted by 0.7) and (2) L_c, Kulback-Leibler to improve the clustering performance.

D. Pronostic of Spitzoid melanocytic neoplasms

As discussed in Section II-A, the database used is also composed of UMP samples. Predicting the prognosis of these tumors is a diagnostic challenge for the pathologists. This is the reason why, in this work, a subspace prototypical algorithm

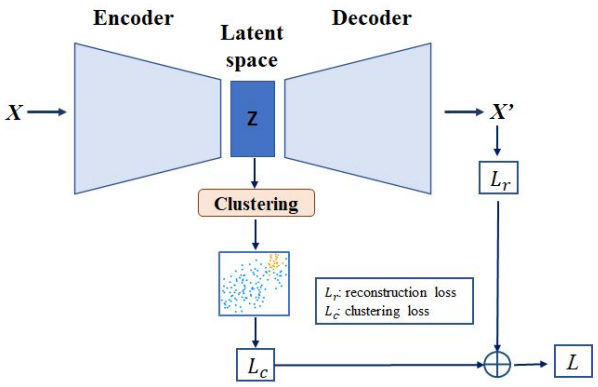

Fig. 3. Deep embedded clustering approach to detect spitzoid melanoma using DNA methylation data. The proposed algorithm is trained minimizing both, reconstruction and clustering loss.

is proposed for predicting the unknown evolution of these patients. These networks are based on the idea that exits a latent space in which the features are grouped around a single prototype representation for each class.

A nonlinear mapping between the DNAm data and a latent space was achieved with the autoencoder specified in Section II-B. After obtaining the embedded space (the encoder output) of the melanoma and nevus samples, the prototype of each class was calculated as follows:

$$\mathbf{p}_k = \frac{1}{|S_k|} \sum_{(x_i, y_i) \in S_k} f_\phi(x_i) \tag{5}$$

where S_k denotes the set of examples labeled with the class k and f_ϕ the function that transforms the input data x_i to latent space features.

The latent space of UMP patients was inferred through the encoder stage of the models trained with malignant and benign samples. Subsequently, the final confidence values for prognosis estimation (i.e. melanoma or nevus) of UMP cases were calculated. In this case, the Euclidean distance was used:

$$d(\mathbf{p}_k, f_\phi(\mathbf{q})) = \sqrt{(p_{k,i} - f_\phi(q_i))^2} \tag{6}$$

where $i = 1, ..., N$. N is the latent space dimension and $\mathbf{q}$ the initial DNAm feature vector of UMP patients. Subsequently, the class probability (melanoma and nevus) was calculated for each query point q_i via the softmax function of distances to prototypes in the embedding space:

$$prob(y = k|\mathbf{q}) = \frac{exp(-d(\mathbf{p}_k, f_\phi(\mathbf{q})))}{\sum_{k'} exp(-d(\mathbf{p}_{k'}, f_\phi(\mathbf{q}))} \tag{7}$$

III. Results and Discussion

In this section, we present the results achieved in each proposed approach. We expose in Table II a comparison of the different dimensionality reduction and classification algorithms. Several figures of merit are calculated to evidence the differences between using supervised or non-supervised classification algorithms. In particular, sensitivity (SN), specificity (SPC), positive predictive value (PPV), negative predictive value (NPV), F1-score (F1S), accuracy (ACC) and area under the ROC curve (AUC) are employed. Note that, to obtain the prediction of all patients with benign and malignant melanocytic neoplasms with supervised classification methods, a weighted K-fold validation was used.

TABLE II
Results obtained for the proposed classification algorithms. For the MLP algorithm, the results obtained over validation subset are shown. However, for the DERC approach, we show the results obtained after the model training.

	AE + MLP	VAE + MLP	DERC
SN	0.75	0.25	**0.88**
SPC	0.92	**1**	0.92
PPV	0.86	**1**	0.88
NPV	0.85	0.67	**0.92**
F1S	0.8	0.40	**0.88**
ACC	0.85	0.65	**0.90**
AUC	0.90	0.70	0.90

Classification results in the latent space of the autoencoder demonstrate SMT diagnosis outperformance with respect to using the embeeding features of the VAE as can be appreciated in Table II. 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. When comparing AE+MLP results with those obtained by the DERC method, it is verified that the latter achieves the best classification results. Therefore, the unsupervised end-to-end training that minimizes the reconstruction data while the refinement of the classification is carried out allows for a better distinction between benign and malignant SMT.

Table III shows the probability and the reliability obtained for the UMP prognostic. As shown in Table III, for most patients designated as uncertain malignant potential, the method predicts a prognosis of benign (63.16%). Therefore, the majority of UMP patients presents features closer to those patients labelled as benign. Fig. 4 represents, in 2D graphics, the distribution of the latent space components (C) of UMP patients with respect to the centroid of each class (malignant and benign).

Thus, with this approach, pathologists have a guideline on the evolution of UMP patients. Since there is no prior work for the Spitzoid classification based on DNAm, a comparison with the state of the art has not been possible.

TABLE III
Prognostic of the UMP patients.

Prognostic	***Samples***	***Probability (%)***	***Reliability*** ($\mu \pm \sigma$)
Melanoma	7	36.84	0.65 ± 0.0511
Benign	12	63.16	0.64 ± 0.0682

IV. Conclusion

In this paper, different deep-learning approaches (supervised and non-supervised) have been presented to elucidate the

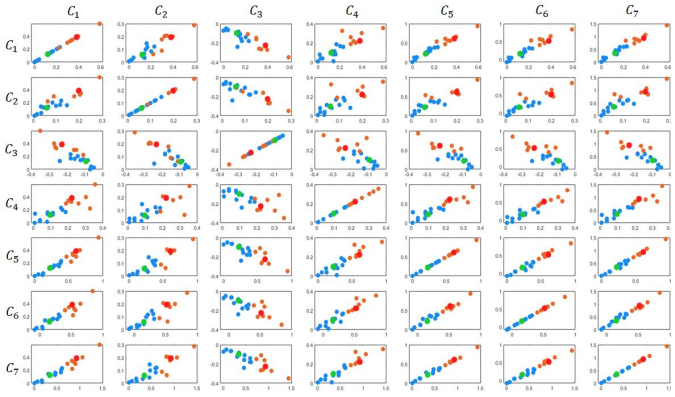

Fig. 4. Latent space representation of UMP patients. Legend: Blue and Orange points represent the patients predicted as benign and malignant, respectively. Green and red color circles are the centroids of each class (benign and malignant).

added value enclosed in the DNAm for Spitzoid neoplasms detection. The reported results suggest the unsupervised deep-clustering-based algorithm as the most promising method. In future research lines, an external validation of the proposed strategy with large databases will be considered.

ACKNOWLEDGMENT

We gratefully acknowledge the support from the Generalitat Valenciana (GVA) with the donation of the DGX A100 used for this work, action co-financed by the European Union through the Operational Program of the European Regional Development Fund of the Comunitat Valenciana 2014-2020 (IDIFEDER/2020/030).

REFERENCES

[1] Z. Apalla, A. Lallas, E. Sotiriou, E. Lazaridou, and D. Ioannides, "Epidemiological trends in skin cancer," *Dermatology practical & conceptual*, vol. 7, no. 2, p. 1, 2017.

[2] U. Leiter, T. Eigentler, and C. Garbe, "Epidemiology of skin cancer," pp. 120–140, 2014.

[3] R. A. DePinho, "The age of cancer," *Nature*, vol. 408, no. 6809, pp. 248–254, 2000.

[4] J. Jaworek-Korjakowska and P. Kłeczek, "Automatic classification of specific melanocytic lesions using artificial intelligence," *BioMed research international*, vol. 2016, 2016.

[5] M. S. Soengas and S. W. Lowe, "Apoptosis and melanoma chemoresistance," *Oncogene*, vol. 22, no. 20, pp. 3138–3151, 2003.

[6] T. Wiesner, H. Kutzner, L. Cerroni, M. C. Mihm Jr, K. J. Busam, and R. Murali, "Genomic aberrations in spitzoid melanocytic tumours and their implications for diagnosis, prognosis and therapy," *Pathology*, vol. 48, no. 2, pp. 113–131, 2016.

[7] ——, "Genomic aberrations in spitzoid melanocytic tumours and their implications for diagnosis, prognosis and therapy," *Pathology*, vol. 48, no. 2, pp. 113–131, 2016.

[8] P. Gerami, S. S. Jewell, L. E. Morrison, B. Blondin, J. Schulz, T. Ruffalo, P. Matushek IV, M. Legator, K. Jacobson, S. R. Dalton *et al.*, "Fluorescence in situ hybridization (fish) as an ancillary diagnostic tool in the diagnosis of melanoma," *The American journal of surgical pathology*, vol. 33, no. 8, pp. 1146–1156, 2009.

[9] K. Conway, S. N. Edmiston, J. S. Parker, P. F. Kuan, Y.-H. Tsai, P. A. Groben, D. C. Zedek, G. A. Scott, E. A. Parrish, H. Hao *et al.*, "Identification of a robust methylation classifier for cutaneous melanoma diagnosis," *Journal of Investigative Dermatology*, vol. 139, no. 6, pp. 1349–1361, 2019.

[10] B. Sánchez-Sendra, C. Martinez-Ciarpaglini, J. F. González-Muñoz, A. Murgui, L. Terrádez, and C. Monteagudo, "Downregulation of intratumoral expression of mir-205, mir-200c and mir-125b in primary human cutaneous melanomas predicts shorter survival," *Scientific reports*, vol. 8, no. 1, pp. 1–14, 2018.

[11] D. Pradhan, G. Jour, D. Milton, V. Vasudevaraja, M. T. Tetzlaff, P. Nagarajan, J. L. Curry, D. Ivan, L. Long, Y. Ding *et al.*, "Aberrant dna methylation predicts melanoma-specific survival in patients with acral melanoma," *Cancers*, vol. 11, no. 12, p. 2031, 2019.

[12] G. Micevic, N. Theodosakis, and M. Bosenberg, "Aberrant dna methylation in melanoma: biomarker and therapeutic opportunities," *Clinical epigenetics*, vol. 9, no. 1, pp. 1–15, 2017.

[13] R. Lazova, N. Pornputtapong, R. Halaban, M. Bosenberg, Y. Bai, H. Chai, and M. Krauthammer, "Spitz nevi and spitzoid melanomas: exome sequencing and comparison with conventional melanocytic nevi and melanomas," *Modern Pathology*, vol. 30, no. 5, pp. 640–649, 2017.

[14] J. A. Tsou, J. A. Hagen, C. L. Carpenter, and I. A. Laird-Offringa, "Dna methylation analysis: a powerful new tool for lung cancer diagnosis," *Oncogene*, vol. 21, no. 35, pp. 5450–5461, 2002.

[15] M. Esteller, "Epigenetics in cancer," *New England Journal of Medicine*, vol. 358, no. 11, pp. 1148–1159, 2008.

[16] N. Jazayeri and H. Sajedi, "Breast cancer diagnosis based on genomic data and extreme learning machine," *SN Applied Sciences*, vol. 2, no. 1, p. 3, 2020.

[17] Z. Si, H. Yu, and Z. Ma, "Learning deep features for dna methylation data analysis," *IEEE Access*, vol. 4, pp. 2732–2737, 2016.

[18] M. Khwaja, M. Kalofonou, and C. Toumazou, "A deep autoencoder system for differentiation of cancer types based on dna methylation state," *arXiv preprint arXiv:1810.01243*, 2018.

[19] Z. Wang and Y. Wang, "Exploring dna methylation data of lung cancer samples with variational autoencoders," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2018, pp. 1286–1289.

[20] E. Min, X. Guo, Q. Liu, G. Zhang, J. Cui, and J. Long, "A survey of clustering with deep learning: From the perspective of network architecture," *IEEE Access*, vol. 6, pp. 39 501–39 514, 2018.

[21] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *International conference on machine learning*, 2016, pp. 478–487.

[22] G. P. Way and C. S. Greene, "Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders," *BioRxiv*, p. 174474, 2017.

[23] d. A. Rocío, C. Adrián, M. Carlos, and N. Valery, "A deep embedded refined clustering approach for breast cancer distinction based on dna methylation," *arXiv preprint arXiv:2102.09563*, 2021.