Metastatic State of Colorectal Cancer can be Accurately Predicted with Methylome

Somayah Albaradei
King Abdullah University of Science and Technology, Computational Bioscience Research Center (CBRC); Computer, Electrical and Mathematical Sciences and Engineering (CEMSE) Division; Thuwal 23955-6900, Kingdom of Saudi Arabia
King Abdulaziz University, Faculty of Computers and Information Systems; Jeddah, Kingdom of Saudi Arabia
somayah.albaradei@kaust.edu.sa

Maha Thafar
King Abdullah University of Science and Technology, Computational Bioscience Research Center (CBRC); Computer, Electrical and Mathematical Sciences and Engineering (CEMSE) Division; Thuwal 23955-6900, Kingdom of Saudi Arabia
Taif University, Faculty of Computers and Information Technology; Taif, Kingdom of Saudi Arabia

Christophe Van Neste
King Abdullah University of Science and Technology, Computational Bioscience Research Center (CBRC); Computer, Electrical and Mathematical Sciences and Engineering (CEMSE) Division; Thuwal 23955-6900, Kingdom of Saudi Arabia
Ghent University, Center for Medical Genetics Ghent (CMGG), Ghent, Belgium

Magbubah Essack
Vladimir B. Bajic
King Abdullah University of Science and Technology, Computational Bioscience Research Center (CBRC); Computer, Electrical and Mathematical Sciences and Engineering (CEMSE) Division; Thuwal 23955-6900, Kingdom of Saudi Arabia

ABSTRACT

Colorectal cancer (CRC) appears to be the third most common cancer as well as the fourth most common cause of cancer deaths in the world. Its most lethal states are when it becomes metastatic. It is of interest to find tests that can quickly and accurately determine if the patient has already developed metastasis. Changes in methylation profiles have been found to be characteristic of cancers at different stages and can therefore be used to develop diagnostic panels. We developed a deep learning (DL) model (Deep2Met) using methylation profiles of patients with CRC to predict if the cancer is in its metastatic state. Results suggest that our method achieves an AUPR and an average F-score of 96.99% and 94.71%, respectively, making Deep2Met potentially useful for diagnostic purposes. The DL model Deep2Met we developed, shows promise in the diagnosis of CRC based on methylation profiles of individual patients.

CCS Concepts

• **Computing methodologies→Supervised learning**

Keywords

Computational biology; Bioinformatics; Metastasis; Colorectal cancer; DNA methylation; Deep learning; Convolutional neural networks

1. INTRODUCTION

Loss of DNA methylation, coined hypomethylation, was the first epigenetic change observed in various human cancers versus healthy tissue [1]-[3]. Cancers typically exhibit this genome-wide loss of DNA methylation, alongside CpG island hypermethylation in the promoters of tumor suppressor genes, which induces transcriptional silencing [4]-[6]. There are many studies that illuminate the role of methylation in cancers [7], [8]. DNA hypomethylation and hypermethylation were more frequently observed in metastatic cancer than in primary cancers [9], [10]. These DNA methylation changes occur at different subsets of the loci [10], which suggests a correlation between metastatic capacity, gene involvement, and DNA methylation. Such epigenetic changes occur early and are frequently manifested in the pathogenesis of colorectal cancer (CRC)[11].

CRC progression is associated with the inactivation of several tumor-suppressor genes such as APC, p53, DCC, and DPC4/SMAD4, and SMAD2 [12], [13] and mismatch repair (MMR) genes such as hMLH1 and hMSH2 [14]. The MMR genes prevent DNA damage by repairing erroneous insertions and deletions during the DNA replication. Thus, inactivation of the MMR genes produces a CRC hypermutable phenotype, that represents an alternative gene inactivation mechanism, with causes microsatellite instability (MSI) and mutations in genes such as TGFβRII and BAX that have several regulatory functions including the induction of apoptosis [15]-[17]. Some of the MMR genes were shown to be hypermethylated. Oster *et al.* identified and validated CpG islands that are frequently hypermethylated in colorectal adenomas and carcinomas[18]. They reported FLI1, ST6GALNAC5, TWIST1, ADHFE1, JAM2, IRF4, CNRIP1, NRG1, and EYA4 are hypermethylated in adenomas and carcinomas, ABHD9, AOX1 and RERG in carcinomas only, and RAMP2, DSC3, and MLH1 in MSI but not MSS carcinomas [18]. They further reported that some of

the adenomas exhibit the hypermethylation changes, suggesting they could be potential early detection biomarkers of CRC.

TCGA[19], [20] is a pan-cancer project to discover molecular aberrations in cancer that are relevant for initiation or progression. Within TCGA, there is available data for RNA, DNA, protein and methylation. These different "omics" types have been used to predict metastasis separately and in combination in different studies, but here we only focus on the latter. Other studies have also focused specifically on methylation data. Vaquero-Garcia *et al.* [21] predicted uveal melanoma metastasis using general clinical characteristics and chromosomal karyotype features. Their machine learning (ML) models correctly predicting up to 85% of test cases. Decock *et al.* [22] profiled the methylome of primary neuroblastoma tumors to discover methylation biomarkers predicting metastasis. In colorectal cancer, Fouad *et al.* [23] analyzed global DNA methylation of blood leukocytes as a proxy for methylation status of the tumor.

Although there have been several computational methods for identifying cancer markers based on DNA methylation [24]-[26], these computational works are still very few, and there is a big room for improvements. Thus, we develop our deep learning (DL) -based method, Deep2Met, for predicting metastatic CRC using useful and relevant features from DNA methylation data. To our knowledge, no other studies utilized data from the TCGA repository and a DL model on DNA methylation beta values to predict if CRC is in a metastatic state.

2. RELATED WORK

The identification of foretelling biological markers for cancer is critical in cancer research. Nowadays, DNA methylation has been proved to be related to cancer stages and prediction. Markers of early stages CRC [27] and other cancers (e.g., breast cancer [28], [29], ovarian cancer [30], and prostate cancer [31]) can be derived based on methylation profiles, achieving high accuracies. Such approaches suggest that using methylation characterization of cancers could be a viable route towards the development of different diagnostic methods for cancers.

The general problem is to identify sufficiently robust markers. Traditional methods for cancer marker detection are not sufficient, as the resultant individual markers are not very efficient and suffer from low accuracy. Recently [32]-[34] there has been an essential emphasis on developing computational methods that help in cancer detection (i.e., cancer markers identification), utilizing information in the DNA methylation data. These methods can be categorized into two groups: (1) tradition statistical-based methods, and (2) ML-based and DL-based methods.

Statistical-based methods have been applied to find out how to distinguish between normal and tumor samples. In [35], the authors used the methylation values of many healthy samples and defined the average healthy methylation distribution. Then, based on the Jensen-Shannon and Kullback-Leibler distances between healthy and tumor samples, they considered possible indicators of cancer.

ML- and DL-based methods boost the analysis and prediction of cancer and have the ability to deal with large and big data. Different ML and DL approaches have been developed to find out how to distinguish between different types of cancers by utilizing the information in DNA methylation data. Work in [26], aimed to recognize cancer types using restricted Boltzmann

machines DL technique (136 samples with 97.06 % accuracy). used Convolutional Neural Networks (CNN) to explore the relationship between methylated versus unmethylated CpG regions (10'000 samples with 92.87% accuracy). Furthermore, that method can be used to classify the cancer types on a genome-wide scale with high accuracy. Another recent study [25] used a different type of DL approaches based on Deep Autoencoder content retrieval algorithm to predict and differentiate between cancer types using the presence of DNA methylation epigenetic patterns (341 samples with 93.33% accuracy). Integration this DL method with several signal processing and statistical techniques allowed more accurate prediction of cancer presence.

3. METHOD

In this study, we developed a DL-based model, Dep2Met, to predict metastatic CRC based on DNA methylation beta values.

3.1 Datasets

We used information from TCGA consortium (http://cancergenome.nih.gov/), which provides medically relevant data for 30+ types of human cancers. DNA methylation data were obtained using the recent version of Infinium HumanMethylation450 BedChip technology. It includes more than 480'000 CpG site probes [36]. The methylation level (beta-value) of each CpG locus is calculated using the following formula:

$$\beta = \frac{max(I_M, 0)}{max(I_M, \ 0) + max(I_U, \ 0) + 100} \qquad (1)$$

3where IM and IU are the intensities of signals of the measured methylated allele and unmethylated allele, respectively; to deal with the case when both IM and IU are small the bias of 100 is added for regularization [37]. We used NCI Genomic Data Commons (GDC) data portal to access the DNA methylation data aligned with the reference human genome GRCh37 (hg19). Clinical information on TCGA based on the American Joint Committee on Cancer (AJCC), provides a "metastasis_pathologic_pm" category, which tells whether there are distant metastases for the sample under investigation or no. In this "label" column M0 means no distant metastasis and M1 means a distant metastasis cancer. About 300 samples were extracted (90 with metastatic CRC data and 211 with non-metastatic CRC data).

3.2 Data Preprocessing

Extracted beta-values that correspond to the ratio of methylation intensities between the methylated probe and the overall methylation intensity (sum of unmethylated and methylated probes) ranges between [0,1]. Under the ideal conditions, if the corresponding CpG site is methylated, the value will be 1, else it will be zero.

Initially, there were 485'577 rows, where each row contains beta-values for a specific CpG site. After we eliminated probes with all null values, we finally retained 379'520 rows. Beta-values for each sample were extracted and converted into a matrix of dimensions (593×640).

Using different random number generation seeds, data were 10 times randomly split into three portions: 70% training (around 210 samples), 15% (around 45 samples) validation, and 15% testing (around 45 samples), as shown in Figure 1. The model

generation and testing have been used for each data split, and the results were averaged.
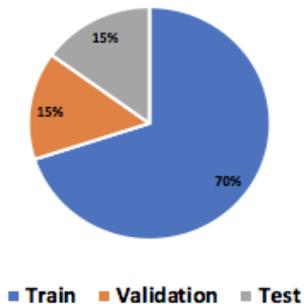

Figure 1. Data split

3.3 Model

The Deep2Met model receives a DNA methylation beta-values as an input and performs feature extraction, feature selection, and final prediction. That is, given DNA methylation beta-values M, our model computes a score f(M) based on a convolutional neural network (CNN), which consists of the following layers: 1) Reconstructed DNA methylation beta-values (section 3.2), 2) horizontal and vertical convolutional layers (CONV), 3) rectified linear unit layer (ReLU), 4) fully connected layer (FC), and 5) SoftMax layer to predict whether the given DNA methylation beta-values correspond to a metastatic or non-metastatic case. The structure of Deep2Met is shown in Figure 2.

The CNN model was constructed to capture distinguishing features that find the relation between DNA methylation patterns and metastatic/non-metastatic samples. Input to the model is a matrix $x \in R^{hw}$, where h is the height, and w is the width of the matrix. CONV layers apply parallel filters: three to capture features horizontally ($f1 \in R^{kw}$) and three to capture features vertically ($f2 \in R^{hk}$), where $k \in \{3,5,7\}$. Each CONV layer is followed by ReLU layer to produce features maps. The generated features maps from horizontal and vertical filters were then flattened and concatenated to each other. The concatenated features are then fed to a fully connected (FC) neural network. The output layer, which has two output neurons with SoftMax activation function, receives input from the FC layer, and perform the prediction. Dropout layer is added to the FC layer to avoid overfitting by randomly removing some neurons during the training of the DL model.

To improve generalization performance on the testing set, we applied an early stopping technique, which monitors the validation error rate and stops the training if the error is not decreasing for ten consecutive epochs. This DL model structure allows our network to learn richer features as DNA methylation beta-values progresses through the network. We used Keras library in R to construct our DL model. The parameters used in this model are shown in Table 1.

Also, as we are working on data containing imbalanced classes, we set the weight of all training samples (Metastatic or non-Metastatic) such that their contributions are balanced when computing the loss. With Keras, we can give weight to the classes by multiplying the loss of each sample by a factor depending on its class. In our model, we set the factor of non-Metastatic sample to be 1, while the factor for Metastatic sample was selected as the ratio number of non-Metastatic samples/number of metastatic samples, which was around 2.2 in our case.
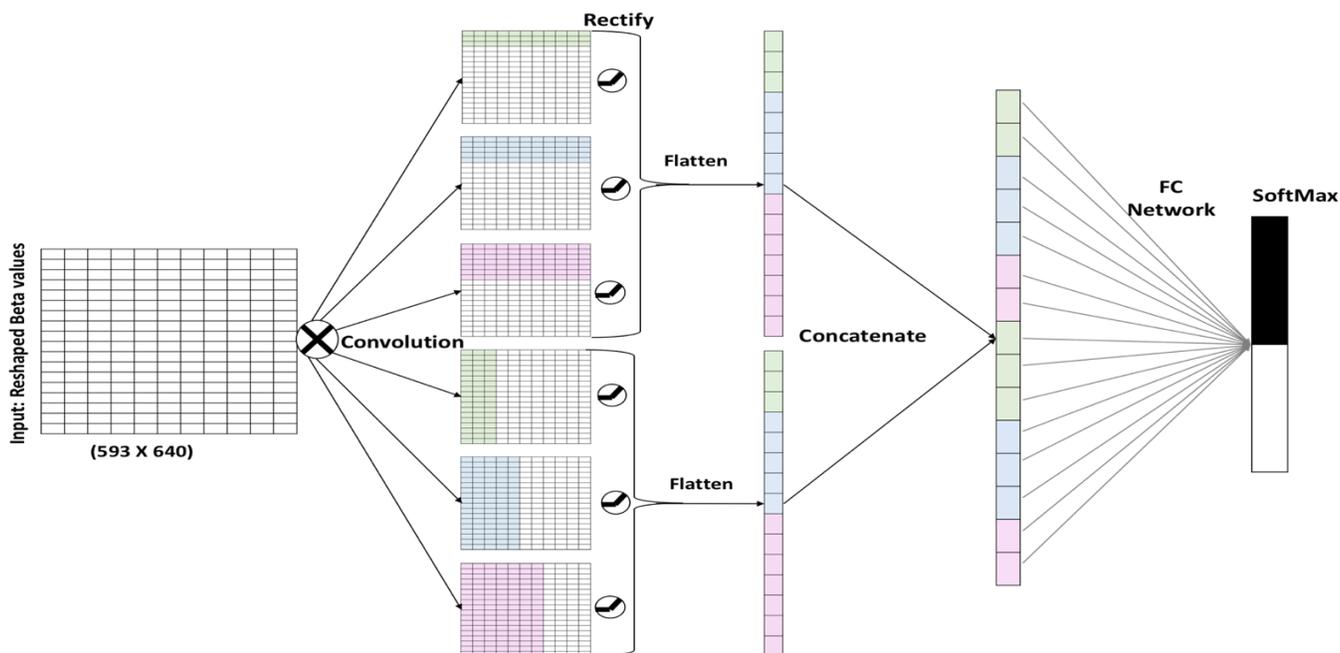

Figure 2. Deep2Met model takes as input a reconstructed DNA methylation beta-values, and apply parallel CONV layers to extract features horizontally and vertically. CONV layers are followed by ReLU layers to produce features maps. Output features are flattened and concatenated feeding into the FC layer followed by a SoftMax layer to predict whether the input corresponds to a metastatic or non-metastatic sample

Table 1 Parameters used in the Deep2Met model

Parameters		Values
Inner Activation function		ReLU
Number of neurons on FC layer		128
Initialization mode		uniform
Batch size		32
Dropout rate		0.5
Optimizer		Adam (learning rate = 0.0001)
Horizontal layers	Number of filters	16
	Filter length	[3 ,5, 7]
	Filter width	h = 593
Vertical layers	Number of filters	16
	Filter length	[3, 5, 7]
	Filter width	w = 640
Output Activation function		SoftMax

Table 2. Evaluation Metrics

Evaluation Metric	Formula	Results
Accuracy	$\dfrac{TP+TN}{TP+FN+TN+}$	96.17%
Sensitivity (Se)	$\dfrac{TP}{TP+FN}$	96.65%
Specificity	$\dfrac{TN}{TN+FP}$	95.83%
Precision (ppv)	$\dfrac{TP}{TP+FP}$	90.44%
Geometric Mean	$\sqrt{ppv \times Se}$	93.49%
F-score	$\dfrac{2 \times TP}{2 \times TP+FP+FI}$	94.71%

Figure 3. Here we depict the mean DNA methylation level for each sample in the Metastatic and non-Metastatic groups
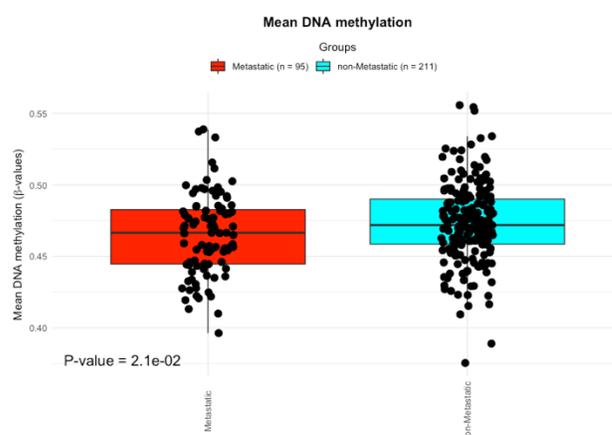
4. RESULTS AND DISCUSSION

Figure 3 demonstrates that it may be possible to develop a predictive model for detecting methylation state for CRC. To evaluate the performance of our model, we computed the statistical measures shown in Table 2. These metrics are calculated based on the prediction of positive (Metastatic) or negative (non-Metastatic) samples. For definition of TP, FN, TN, FP see the Terminology section.

While these metrics are essential to evaluate the model performance, the area under the precision-recall curve (AUPR) plays an even more critical role when estimating performance based on data with imbalanced classes. The AUPR of our model on the test set is **96.99 %**. All reported results are the average performance obtained from 10 random split tests.

To show the effectiveness of such architecture we first defined the base-line architecture. This base-line architecture takes the input as the raw methylation profile (379520 beta-values), followed by a stack of three dense layers each with 250 nodes, followed by an output layer. The AUPR for such a model is 85.64 %. Thus, the proposed Deep2Met model has a promising architecture probably due to the fact that it better utilizes information from the methylation data.

5. CONCLUSIONS

We developed a deep learning model, Deep2Met, that aims at predicting if a patient with colorectal cancer has already entered metastatic state or not. Deep2Met uses methylation profile of the sample and generates the prediction. Although this is a preliminary development, high accuracy of the model prediction makes it potentially useful for diagnosis purposes, though further improvements are necessary.

5.1 Terminology

A True Positive (TP) is the number of metastatic samples that are correctly predicted as positive, while a False Positive (FP) is the number of non-metastatic samples that are wrongly predicted as positive. True Negatives (TN) is the number of non-metastatic samples that are correctly predicted as negative and False Negatives (FN) is the number of metastatic samples that are wrongly predicted as negative.

6. ACKNOWLEDGMENT

7. REFERENCES

[1] Gama-Sosa, M. A., Slagel, V. A., Trewyn, R. W., Oxenhandler, R., Kuo, K. C., Gehrke, C. W. and Ehrlich, M. The 5-methylcytosine content of DNA from human tumors. *Nucleic acids research*, 11, 19 (1983), 6883-6894.

[2] Feinberg, A. P. and Vogelstein, B. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature*, 301, 5895 (1983), 89.

[3] Aryee, M. J., Wu, Z., Ladd-Acosta, C., Herb, B., Feinberg, A. P., Yegnasubramanian, S. and Irizarry, R. A. Accurate genome-scale percentage DNA methylation estimates from microarray data. *Biostatistics*, 12, 2 (2010), 197-210.

[4] Rauch, T. A., Zhong, X., Wu, X., Wang, M., Kernstine, K. H., Wang, Z., Riggs, A. D. and Pfeifer, G. P. High-resolution mapping of DNA hypermethylation and hypomethylation in lung cancer. *Proceedings of the National Academy of Sciences*, 105, 1 (2008), 252-257.

[5] Cho, N. Y., Kim, B. H., Choi, M., Yoo, E., Moon, K. C., Cho, Y. M., Kim, D. and Kang, G. Hypermethylation of CpG island loci and hypomethylation of LINE-1 and Alu repeats in prostate adenocarcinoma and their relationship to clinicopathological features. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland*, 211, 3 (2007), 269-277.

[6] Frigola, J., Solé, X., Paz, M. F., Moreno, V., Esteller, M., Capellà, G. and Peinado, M. A. Differential DNA hypermethylation and hypomethylation signatures in colorectal cancer. *Human molecular genetics*, 14, 2 (2004), 319-326.

[7] Yang, X., Gao, L. and Zhang, S. Comparative pan-cancer DNA methylation analysis reveals cancer common and specific patterns. *Briefings in bioinformatics*, 18, 5 (2016), 761-773.

[8] Youn, A., Kim, K. I., Rabadan, R., Tycko, B., Shen, Y. and Wang, S. A pan-cancer analysis of driver gene mutations, DNA methylation and gene expressions reveals that chromatin remodeling is a major mechanism inducing global changes in cancer epigenomes. *BMC medical genomics*, 11, 1 (2018), 98.

[9] Bedford, M. T. and Van Helden, P. D. Hypomethylation of DNA in pathological conditions of the human prostate. *Cancer research*, 47, 20 (1987), 5274-5276.

[10] Smiraglia, D., Smith, L., Lang, J., Rush, L., Dai, Z., Schuller, D. and Plass, C. Differential targets of CpG island hypermethylation in primary and metastatic head and neck squamous cell carcinoma (HNSCC). *Journal of medical genetics*, 40, 1 (2003), 25-33.

[11] Okugawa, Y., Grady, W. M. and Goel, A. Epigenetic alterations in colorectal cancer: emerging biomarkers. *Gastroenterology*, 149, 5 (2015), 1204-1225. e1212.

[12] Kinzler, K. W. and Vogelstein, B. Lessons from hereditary colorectal cancer. *Cell*, 87, 2 (1996), 159-170.

[13] White, R. L. Tumor suppressing pathways. *Cell*, 92, 5 (1998), 591-592.

[14] Peltomäki, P. and de la Chapelle, A. *Mutations predisposing to hereditary nonpolyposis colorectal cancer*. Elsevier, City, 1997.

[15] Yashiro, M., Hirakawa, K. and Boland, C. R. Mutations in TGFbeta-RII and BAX mediate tumor progression in the later stages of colorectal cancer with microsatellite instability. *BMC cancer*, 10, 1 (2010), 303.

[16] Westphal, D., Kluck, R. and Dewson, G. Building blocks of the apoptotic pore: how Bax and Bak are activated and oligomerize during apoptosis. *Cell death and differentiation*, 21, 2 (2014), 196.

[17] Carberry, S., D'Orsi, B., Monsefi, N., Salvucci, M., Bacon, O., Fay, J., Rehm, M., McNamara, D., Kay, E. W. and Prehn, J. H. The BAX/BAK-like protein BOK is a prognostic marker in colorectal cancer. *Cell death & disease*, 9, 2 (2018), 125.

[18] Øster, B., Thorsen, K., Lamy, P., Wojdacz, T. K., Hansen, L. L., Birkenkamp-Demtröder, K., Sørensen, K. D., Laurberg, S., Ørntoft, T. F. and Andersen, C. L. Identification and validation of highly frequent CpG island hypermethylation in colorectal adenomas and carcinomas. *International journal of cancer*, 129, 12 (2011), 2855-2866.

[19] Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J. M. and Network, C. G. A. R. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45, 10 (2013), 1113.

[20] Tomczak, K., Czerwińska, P. and Wiznerowicz, M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary oncology*, 19, 1A (2015), A68.

[21] Vaquero-Garcia, J., Lalonde, E., Ewens, K. G., Ebrahimzadeh, J., Richard-Yutz, J., Shields, C. L., Barrera, A., Green, C. J., Barash, Y. and Ganguly, A. PRiMeUM: a model for predicting risk of metastasis in uveal melanoma. *Investigative ophthalmology & visual science*, 58, 10 (2017), 4096-4105.

[22] Decock, A., Ongenaert, M., Van Criekinge, W., Speleman, F. and Vandesompele, J. DNA methylation profiling of primary neuroblastoma tumors using methyl-CpG-binding domain sequencing. *Scientific data*, 3 (2016), 160004.

[23] Fouad, M. A., Salem, S. E., Hussein, M. M., Zekri, A. R. N., Hafez, H. F., Eldesouky, E. E. and Shouman, S. A. Impact of Global DNA Methylation in Treatment Outcome of Colorectal Cancer Patients. *Frontiers in pharmacology*, 9 (2018), 1173.

[24] Chatterjee, S., Iyer, A., Avva, S., Kollara, A. and Sankarasubbu, M. Convolutional Neural Networks In Classifying Cancer Through DNA Methylation. *arXiv preprint arXiv:1807.09617* (2018).

[25] Khwaja, M., Kalofonou, M. and Toumazou, C. A Deep Autoencoder System for Differentiation of Cancer Types Based on DNA Methylation State. *arXiv preprint arXiv:1810.01243* (2018).

[26] Si, Z., Yu, H. and Ma, Z. Learning deep features for dna methylation data analysis. *IEEE Access*, 4 (2016), 2732-2737.

[27] Hicham Mansour, R. I., Vladimir Bajic *Composite Biomarkers For Non-Invasive Screening, Diagnosis And Prognosis Of Colorectal Cancer*. City, 2018.

[28] Hicham Mansour, R. I., Vladimir Bajic *Methylation Biomarkers for Breast Cancer*. City, 2018.

[29] Vladimir Bajic, H. M., Roberto Incitti *Methylation Biomarkers for Breast Cancer*. City, 2018.

[30] Hicham Mansour, R. I., Vladimir Bajic *Methylation Biomarkers for Ovarian Cancer*. City, 2018.

[31] Vladimir Bajic, H. M., Roberto Incitti *Methylation Biomarkers for Prostate Cancer*. City.

[32] Kel, A., Boyarskikh, U., Stegmaier, P., Leskov, L. S., Sokolov, A. V., Yevshin, I., Mandrik, N., Stelmashenko, D., Koschmann, J. and Kel-Margoulis, O. Walking pathways with positive feedback loops reveal DNA methylation biomarkers of colorectal cancer. *BMC bioinformatics*, 20, 4 (2019), 119.

[33] Peng, H., Wang, S., Pang, L., Yang, L., Chen, Y. and Cui, X.-b. Comprehensive bioinformation analysis of methylated and differentially expressed genes in esophageal squamous cell carcinoma. *Molecular omics*, 15, 1 (2019), 88-100.

[34] Lin, C., Yuan, G., Hu, Z., Zeng, Y., Qiu, X., Yu, H. and He, S. Bioinformatics analysis of the interactions among lncRNA, miRNA and mRNA expression, genetic mutations and epigenetic modifications in hepatocellular carcinoma. *Molecular medicine reports*, 19, 2 (2019), 1356-1364.

[35] Ramakrishnan, N. and Bose, R. Analysis of healthy and tumour DNA methylation distributions in kidney-renal-clear-cell-carcinoma using Kullback–Leibler and Jensen–Shannon distance measures. *IET systems biology*, 11, 3 (2017), 99-104.

[36] Zhang, X., Mu, W. and Zhang, W. On the analysis of the illumina 450k array data: probes ambiguously mapped to the human genome. *Frontiers in genetics*, 3 (2012), 73.

[37] Du, P., Zhang, X., Huang, C.-C., Jafari, N., Kibbe, W. A., Hou, L. and Lin, S. M. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC bioinformatics*, 11, 1 (2010), 587..