Feature Selection and Extraction Framework for DNA Methylation in Cancer

Abeer A. Raweh
Dept. of Computer Science,
Faculty of Computers and
Information, Cairo University
Cairo, Egypt

Mohammad Nassef
Dept. of Computer Science,
Faculty of Computers and
Information, Cairo University
Cairo, Egypt

Amr Badr
Dept. of Computer Science,
Faculty of Computers and
Information, Cairo University
Cairo, Egypt

***Abstract* —Feature selection methods for cancer classification are aimed to overcome the high dimensionality of the biomedical data which is a challenging task. Most of the feature selection methods based on DNA methylation are time consuming during testing phase to identify the best pertinent features subset that are relevant to accurate prediction. However, the hybridization between feature selection and extraction methods will bring a method that is far fast than only feature selection method. This paper proposes a framework based on both novel feature selection methods that employ statistical variation, standard deviation and entropy, along with extraction methods to predict cancer using three new features, namely, Hypomethylation, Midmethylation and Hypermethylation. These new features represent the average methylation density of the corresponding three regions. The three features are extracted from the selected features based on the analysis of the methylation behavior. The effectiveness of the proposed framework is evaluated by the breast cancer classification accuracy. The results give 98.85% accuracy using only three features out of 485,577 features. This result proves the capability of the proposed approach for breast cancer diagnosis and confirms that feature selection and extraction methods are critical for practical implementation.**

Keywords—DNA methylation, feature selection; feature extraction; cancer classification; epigenetics; biomarkers; hypomethylation; hypermethylation; methylation

I. INTRODUCTION

Cancer is a leading cause of death worldwide, it begins when some cells in a part of the body start to grow out of control. Despite the presence of more than one type of cancer that differ in the way of growing cells and spreading, the development of all these kinds is driven by "genetic alterations" and "epigenetic changes" of the DNA genome [1]. Recent research increases evidences that the epigenetic modifications play a critical role in human cancer. These modifications are heritable changes in a cellular phenotype that are independent of alterations in the DNA sequence [2], [3]. Many studies of epigenetic aberrations in tumors prove that the biology of DNA methylation is the most potential epigenetic marker for cancer detection in spite of many other epigenetic alterations in the mammalian genome such as post-translational modifications of histones, chromatin remodeling and microRNAs patterns [4]. Actually, DNA methylation acts as a gene-silencing mechanism to turn off specific genes due to its significant effects on gene expressions and the architecture of the nucleus of the cell [5]. Chemically, DNA methylation is a relatively stable chemical modification resulting from the addition of a methyl (CH3) group at the carbon 5 position of the cytosine or guanine nucleotides in the context of 5'-CG-3' (CpG dinucleotide) by DNA methyltransferase (DNMT) enzymes [6]. Not all CpG sites in the genome are methylated; CpG islands "regions that are containing a high frequency of CpG dinucleotides" are usually not methylated in normal cells [7]. Throughout the genome, there are two types of cancer-associated DNA methylation based on the methylation level called hypermethylation and hypomethylation. Hypermethylation "the methylation exceeds normal methylation level" of tumor suppressor gene affecting the gene expression and proteins involved in cancer manifestation. On the other hand, hypomethylation "the methylation beneath normal methylation level" has been observed frequently in solid tumors [8].

Due to the huge number of probes in the DNA, the importance of providing researchers and scientists with novel, accurate and robust computational tools for studying the whole genome for the cancer predication is widely increasing. Most of the probes of the mammalian tumors genome are irrelevant classification factors and may have bad effect by introducing noises and hence decreasing predication accuracy [9]. Ideally, a good dimensionality reduction method should eliminate these irrelevant probes while at the same time retain all the highly discriminative probes. Therefore, using feature selection and extraction techniques in cancer predication becomes essential to identify the informative probes that underlie the pathogenesis of tumor cell proliferation. Thus, many recent researches applied feature selection and extraction techniques to extract useful information and diagnosis the tumor [10]-[15].

In this paper, we propose a framework based on feature selection and extraction methods, to rid of irrelevant information and improve cancer classification accuracy based on DNA methylation data. First, a novel feature selection based on statistical variation and standard deviation is utilized for identifying the small set of discriminative methylated DNA probes, afterwards, the average methylation density of three regions (hypomethylation, midmethylation and hypermethylation) is calculated as new extracted features to predict cancer.

The reminder of this paper is organized as follows. Section II elaborates on previous work, Section III presents

the attempted dataset and proposed framework, Section IV discusses our experimental results and the last Section V contains concluding remarks and demonstrates future work.

II. Related Works

To increase the accuracy and handle the dramatically increasing tumor feature data and information, a number of researchers have turned to feature selection and extraction techniques for predicting cancer. Feature selection (FS) is one of the important steps in classification modeling of cancer based on DNA methylation [16], it could be used for eliminating unnecessary information to reduce the high dimensionality of the data. Whereas feature extraction also called data transformation, is the process of transforming the feature data into a quantified data type instead of recognizing new patterns to represent the data.

In the past decade, many feature selection and extraction methods have been proposed, resulting in great improvements of classification. Li *et al.* [10] proposed a gene extraction method by using two standard feature extraction methods, namely, the T-test method and kernel partial least squares (KPLS) in tandem. Zheng *et al.* [11] developed a hybrid of K-means and support vector machine (K-SVM) algorithms to diagnosis breast cancer disease. Kopriva *et al.* [12] proposed a general feature extraction method for cancer prediction based on the linear transformation constructed by tensor decomposition. A novel method using wavelet analysis, genetic algorithm, and Bayes classifier proposed by Liu *et al.* [13] was applied to detect the prognostic biomarkers of survival in colorectal cancer patients. Fontes *et al.* [14] applied feature extraction techniques such as *F-score, p-value rank* and *wrapper approaches* in order to identify which probes presented higher significance in breast cancer prediction. D.L. Tong [15] proposed an innovative hybridized model based on genetic algorithms (GAs) and artificial neural networks (ANNs), to extract the highly differentially expressed genes for specific cancer pathology. Anuradha *et al.* [17] gave a comparative study to identify the best feature extraction technique to classify Oral cancers. Zhuang *et al.* [16] performed another good comparison study of feature selection and classification methods in DNA using the Illumina Infinium platform. Cai *et al.* [18] used Ensemble-based feature extraction methods to capture the unbiased, informative as well as compact molecular signatures followed by SVM trained with Incremental Feature Selection (IFS) strategy to predict subtypes of lung cancer. A novel multiclass feature selection and classification system proposed by Sebastian *et al*. [19] for data merged from different molecular biomedical techniques demonstrated that the feature selection step is crucial in high dimension data classification problems. Furthermore, Baur *et al.* [20] developed a feature selection algorithm based on sequential forward selection to compute gene centric DNA methylation using probe level DNA methylation data. Valavanis *et al*. [8] used semantics information included in the Gene Ontology (GO) tree by graph-theoretic methodology in order to select cancer epigenetic biomarkers.

III. Proposed Framework

A. Dataset

In this study, we conducted experiments on a dataset of large collection of cancer methylomes obtained from The Cancer Genome Atlas (TCGA) using the Human Infinium 450k assay for 4034 cancer and normal tissue samples. The dataset was downloaded from Max Planck Institute for Informatics (MPI) with a software tool for large-scale analysis that yields detailed hypertext reports and interpretation of the DNA methylation data "RnBeads" [21]. As listed in Table 1, the dataset contains several types of cancer: blood, breast, intestinal, brain and other types of cancer. The degree of DNA methylation that extracted from the regions: 31195 promoters, 31033 genes, 485577 probes and 26662 CpG Islands quantified numerically as values.

B. Proposed Framework

The proposed framework is made for detecting cancer based on methylated DNA probes, there are three main steps to be followed in this framework. These steps are feature selection, feature extraction and classification. Fig. 1 shows the architecture of the proposed framework.

C. Feature Selection Methods

Feature selection methods in cancer classification issues are aimed at identifying the minimal-sized subset of markers that are relevant to accurate prediction. To achieve this target, we propose two novel feature selection methods. The first one uses statistical variation in terms of standard deviation in order to select the most informative probes which distinguish normal tissue from cancer. This method measures the differences of probe methylation in all samples compared with the dispersion of this probe methylation in each class (Normal, Cancer) separately. Thus, the discriminative value (*DV*) according to the proposed feature selection for each probe (*X*) based on DNA methylation as an input is defined as:

$$\boldsymbol{DV1}(\boldsymbol{X}) = \frac{\frac{\sum(x-\bar{x})^2}{n-1}}{\sqrt{\frac{\sum(x^+-\bar{x}^+)^2}{n^+-1}}+\sqrt{\frac{\sum(x^--\bar{x}^-)^2}{n^--1}}} \quad (1)$$

TABLE I. Cancer Types in the Attempted Dataset

Cancer Type	No. of Normal Samples	No. of Tumor Samples
Breast invasive carcinoma	98	573
Colon adenocarcinoma	38	253
Glioblastoma multiforme	1	125
Head and Neck squamous cell carcinoma	50	373
Kidney renal clear cell carcinoma	160	283
Acute Myeloid Leukemia	0	194
Lung adenocarcinoma	32	409
Lung squamous cell carcinoma	42	360
Rectum adenocarcinoma	7	96
Thyroid carcinoma	56	435
Uterine Corpus Endometrioid Carcinoma	46	393

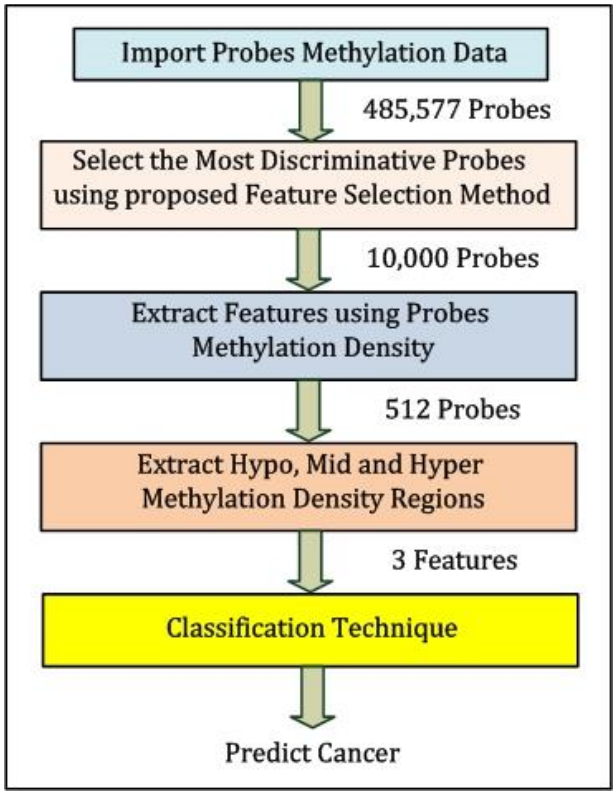

Fig. 1. Architecture of the proposed framework.

Where:

$\overline{x}$ is the methylation average of entire dataset samples for probe *(X)*.

$\overline{x}^+, \overline{x}^-$ are the methylation average of the cancer and normal samples respectively for probe *(X)*.

x is the methylation of entire dataset samples for probe *(X)*.

x^+, x^- are the methylation of the cancer and normal samples respectively for probe *(X)*.

n is the number of all samples.

n^+, n^- are the number of cancer and normal samples respectively.

The second feature selection method is proposed to find the more variational features with less amount of uncertainty involved in its values (less disorder features). The key measure in information theory for measuring uncertainty is the "entropy" which is defined by Claude E. Shannon [22], [23] and considered as a measure to rank features. Regard to this, the above formula *DV1(X)* with entropy is defined as:

$$DV2(X) = \frac{DV1(X)}{H(Y|X)} \quad (2)$$

Where:

$H(Y|X)$ is the entropy for two variables *X* and *Y* that measures the uncertainty of *Y* when *X* is known.

$$H(Y|X) = -\sum p(x) \sum p(y|x) \log_2(p(y|x)) \quad (3)$$

Where:

Y denotes all available classes (Normal and Cancer).

X is the methylation of gene promoter.

$p(x)$ is the probability of interval x

$p(y|x)$ is the probability of class y given interval x.

From 485,577 probes, 10,000 probes are selected using the proposed feature selection methods.

D. Feature Extraction Method

The most discriminative probes (i.e. 10,000 probes) are selected using the proposed feature selection *DV1(X)*. Then these features are extracted using feature extraction methods. Feature extraction is the process which involves for clarifying and detecting the methylation patterns or methylation behavior in the selected probes. As a first step, we use kernel density estimator method [24]; which infers population probability density function of the selected probes; as a feature extraction method, in order to extract 512 features for each sample from the selected 10,000 probes. The kernel density estimate of f at the point x is given by

$$fh(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x-xi}{h}\right) \quad (4)$$

Where *K* denotes to so-called Gaussian kernel function that integrates to one and has mean zero. It defined as:

$$K(u) = \frac{1}{\sqrt{2\prod}} \exp\left(-\frac{1}{2} u^2\right) \quad (5)$$

And h denotes to a smoothing parameter >0 called the bandwidth. The optimal bandwidth that gives better results can be obtained by

$$h_{opt} = \frac{0.9 \text{ X } \partial}{\sqrt[5]{N}} \quad (6)$$

Where, $\partial = \min(\partial, \frac{IQR}{1.34})$ and IQR is the interquartile range that measures the difference between the 75th percentile ($Q3$) and the 25th percentile ($Q1$): $IQR = Q3 - Q1$.

In the second step, for each sample we extract three features from 512 features of kernel density method that have been obtained. The extracted three features are belonging to average methylation density of three regions: Hypomethylation, Middle-methylation (Midmethylation) and Hypermethylation region.

E. Classification

To evaluate the ability of the proposed framework for cancer classification based on methylated probes, the following classifiers: Naïve Bayes, Random Forest, Hoeffiding Tree, SVM and Simple Logistic were used. The accuracy, F-Measure, Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) of each classifier were used as a metrics for evaluation. 250 samples from breast tissue were used as training data and 348 samples were used as testing data. Furthermore, different approaches were used to study classifier's ability in cancer prediction, where the first experiment used the methylation density of whole probes (485,577 probes), the second experiment used methylation density of most discriminative probes chosen by *DV1(X)* (10,000 probes) and the last experiment used three features only "average methylation density of three regions (Hypo, Mid, Hyper methylation)". The next section shows the testing accuracy, F-Measure, Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) of each machine learning

technique. Through these experiments, the reader can observe the ability of classifier in cancer prediction using only the extracted three features.

IV. Results Analysis and Discussion

Firstly, this section compares the proposed feature selection methods, *DV1(X)* and *DV2(X)*, with the existing feature selection methods such as: F-Score, Chi-Squared, Information Gain, and Symmetrical Uncertainty (SU) to evaluate their ability to select the most discriminative probes for cancer classification. To ensure a fair comparison, we conduct the experiments on breast tissue which contains the maximum number of samples in the dataset as illustrated in Table 1. For the breast tissue dataset, 250 samples were used as training data whereas 348 samples were used as testing data. Tables 2 to 4 reports the testing accuracies, F-Measure, Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) of some machine learning techniques such as: Naïve Bayes, Random Forest, Hoeffiding Tree, SVM and Simple Logistic for 31 selected probes. The results show that the proposed methods, *DV1(X)* and *DV2(X)*, always outperform the existing feature selection methods in terms of the predication accuracy.

TABLE II. Prediction Accuracy of Different Classifiers based on Different Feature Selection Methods

Classification Techniques / FS Methods	**Naïve Bayes**	**Random Forest**	**Hoeffiding Tree**	**SVM**	**Simple Logistic**
Proposed Method *DV1(X)*	98.85%	99.14%	98.85%	98.85%	98.56%
Proposed Method *DV2(X)*	99.43%	99.14%	99.43%	99.43%	98.28%
F-Score	98.28%	98.56%	98.28%	97.7%	97.7%
Chi- Squared	98.28%	98.85%	98.28%	98.56%	97.13%
Information Gain	97.99%	98.28%	97.99%	98.85%	96.84%
SU	98.85%	98.85%	98.85%	98.28%	96.55%

TABLE III. F-Measure of Different Classifiers based on Different Feature Selection Methods

Classification Techniques / FS Methods	**Naïve Bayes**	**Random Forest**	**Hoeffiding Tree**	**SVM**	**Simple Logistic**
Proposed Method *DV1(X)*	95.9%	96.8%	95.9%	95.9%	94.8%
Proposed Method *DV2(X)*	97.9%	96.9%	97.9%	97.9%	93.3%
F-Score	93.6%	94.8%	93.6%	91.7%	91.7%
Chi- Squared	94%	95.9%	94%	94.9%	90.4%
Information Gain	93.1%	93.5%	93.1%	95.8%	88.4%
SU	95.9%	95.7%	95.9%	94%	88.7%

TABLE IV. Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) of Different Classifiers based on Different Feature Selection Methods

Classification Techniques / FS Methods		**Naïve Bayes**	**Random Forest**	**Hoeffiding Tree**	**SVM**	**Simple Logistic**
Proposed Method DV1(X)	MAE	0.01	0.04	0.01	0.01	0.02
	RMSE	0.1	0.1	0.1	0.1	0.09
Proposed Method DV2(X)	MAE	0.005	0.03	0.005	0.005	0.02
	RMSE	0.07	0.1	0.07	0.07	0.12
F-Score	MAE	0.01	0.03	0.01	0.02	0.06
	RMSE	0.13	0.11	0.13	0.15	0.14
Chi-Squared	MAE	0.01	0.03	0.01	0.01	0.1
	RMSE	0.13	0.1	0.13	0.11	0.15
Information Gain	MAE	0.01	0.05	0.01	0.01	0.04
	RMSE	0.13	0.13	0.13	0.1	0.14
SU	MAE	0.01	0.03	0.01	0.01	0.19
	RMSE	0.1	0.1	0.1	0.13	0.22

Furthermore, to demonstrate the ability of the proposed framework for cancer classification based on methylated probes, the following Tables 5 to 7 reports the testing accuracy, F-Measure, Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) of different machine learning techniques. These tables compares the results of three approaches: the first one when using the whole probes density, the second one when using the density of 10,000 Probes choosing by *DV1(X)*, and the third one when using the three extracted features (average density of Hypo, Mid and Hyper regions). The results prove the capability of the proposed approach in cancer prediction using only three extracted features.

In addition, this section makes an analysis and comparison of the behavior of the valuable data in probe regions "DNA methylation" in breast tissue samples (normal and cancer). Fig. 2 shows the average methylation of 98 normal samples and 500 cancer samples in the whole probes "485577 probes".

TABLE V. Comparison of Accuracy Obtained by Different Classifiers based on Different Approaches

Classifier / Approach	**Naïve Bayes**	**Random Forest**	**Hoeffiding Tree**	**SVM**	**Simple Logistic**
Whole Probes Density	80.17%	87.36%	79.89%	83.05%	82.76%
Density of 10,000 Probes choosing by *DV1(X)*	98.56%	97.70%	98.56%	97.70%	97.70%
Average density of Hypo,Mid and Hyper regions	98.28%	98.56%	98.28%	98.85%	98.28%

TABLE VI. F-MEASURE OBTAINED BY DIFFERENT CLASSIFIERS BASED ON DIFFERENT APPROACHES

Classifier / Approach	Naïve Bayes	Random Forest	Hoeffding Tree	SVM	Simple Logistic
Whole Probes Density	47.3%	42.1%	47%	27.2%	25%
Density of 10,000 Probes choosing by *DV1(X)*	94.6%	91.5%	94.6%	91.5%	91.1%
Average density of Hypo,Mid and Hyper regions	93.5%	94.6%	93.5%	95.7%	93.3%

TABLE VII. MEAN ABSOLUTE ERROR (MAE) AND ROOT MEAN SQUARED ERROR (RMSE) OBTAINED BY DIFFERENT CLASSIFIERS BASED ON DIFFERENT APPROACHES

Approach / Classifier	Whole Probes Density		Density of 10,000 Probes choosing by *DV1(X)*		Average density of Hypo,Mid and Hyper regions	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
Naïve Bayes	0.19	0.44	0.01	0.11	0.01	0.12
Random Forest	0.17	0.29	0.03	0.13	0.03	0.13
Hoeffiding Tree	0.2	0.44	0.01	0.11	0.01	0.12
SVM	0.16	0.41	0.02	0.15	0.01	0.1
Simple Logistic	0.21	0.35	0.04	0.13	0.04	0.13

It is clear that the methylation behavior can be divided into three regions: low level of methylation region "hypomethylation", middle level of methylation region "midmethylation" and high level of methylation region "hypermethylation". This figure demonstrates that there is a difference between methylation behavior in normal and cancer samples, where the density of methylation in normal samples are lower in cancer samples. This difference, however, is not totally clear. Moreover, as shown in Table 5, the Random Forest classifier gave 87.36% as a higher prediction accuracy using the density of whole probes approach.

For a deep dive into the difference between methylation behavior in normal and cancer samples, we concentrated on the most informative probes that are relevant to accurate cancer predication. Fig. 3 shows the average methylation of the most discriminative probes (10,000 Probes choosing by *DV1(X)* in all normal and cancer samples. As shown in this figure, the difference is more clearly, where the density of hypomethylation and hypermethylation are lower in cancer samples. The decreasing in density of hypomathylation in the cancer sample means that, the amount of methylation is increased in these regions, and thus all the respective genes are turned from active genes to silent genes. By contrast, the decreasing density of hypermathylation in a cancer sample means decreasing amount of methylation; therefore all the respective genes in these regions are turned from silent genes to active genes. Furthermore, using the density of discriminative probes "10,000 probes" in cancer prediction improves classifier accuracy, where both Naïve Base and Hoeffiding Tree classifier gave 98.56% as a higher prediction accuracy using this approach. Moreover, Fig. 4 compares the behavior of methylation in cancer cell in some other tissues such as: Colon, Kidney and Uterine.

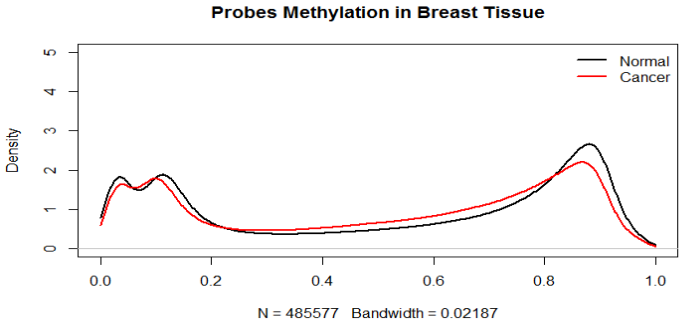

Fig. 2. Whole probes methylation in Breast tissue.

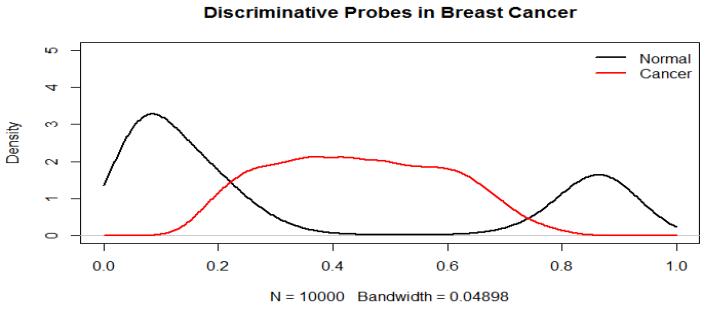

Fig. 3. Methylation of the most discriminative probes (10,000 Probes choosing by *DV1(X)*).

We found that the behavior of methylation is the same in all tissues, increasing methylation of hypomethylation and decreasing methylation of hypermethylation.

As we mentioned in our experiments, we extracted three features from 512 features of kernel density estimator method. These three features belong to average methylation density of three regions: hypomethylation, midmethylation and hypermethylation region. To obtain these features, we calculated the intersection points between normal and cancer curve. As shown in Fig. 5, *0.223092* and *0.741683* are intersection points between the curves, and thus, the curves can be divided into hypomethylation, midmethylation and finally hypermethylation region. Fig. 5 shows the intersection points and these three regions, where letter *A* denotes to hypomethylation region, letter *B* denotes to midmethylation region and letter *C* denotes to hypermethylation region. In addition, as shown in Table 5, using these three features out of 485577 features "probes" in cancer predication improves classifier accuracy (from 83.05% to 98.85%), for SVM classifier which gave a higher accuracy using this approach. These results emphasize the capability of our proposed framework in cancer classification and illustrate the importance of using feature selection and extraction for accurate cancer predication.

To provide a better understanding of the DNA methylation mechanism that plays a major role in the development and progression of cancer, we analyze the top 31 probes that have been generated from the proposed feature selection methods (*DV1* and *DV2*) and used in the classification experiments.

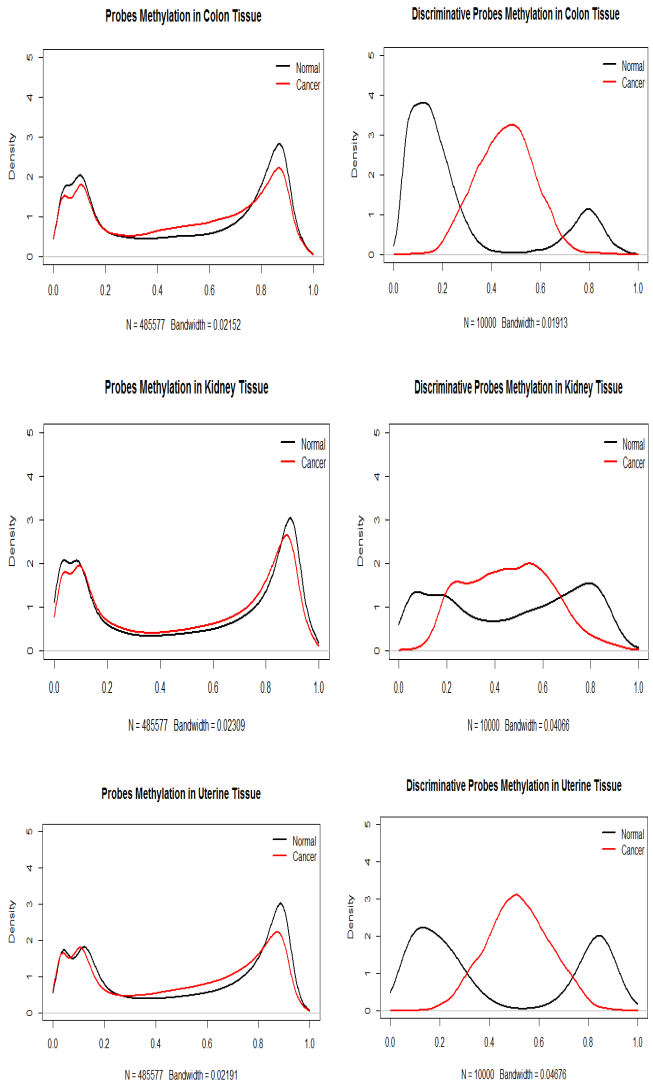

Fig. 4. Methylation behavior in Colon, Kidney and Uterine tissues.

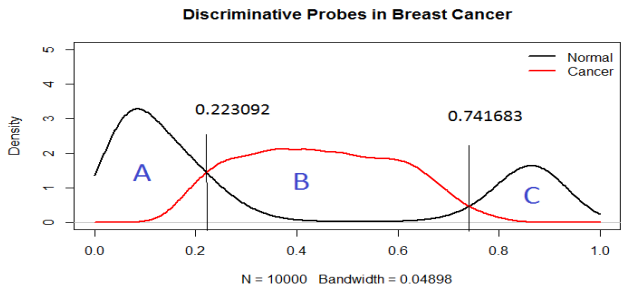

Fig. 5. Intersection points and Hypomethylation, Midmethylation, Hypermethylation regions.

Therefore, we confirm that the role of DNA methylation is to activate or silence some genes by decreasing or increasing their methylation respectively. Furthermore, we examine the ability of a new subset of probes to predict cancer, the subset contains common probes from the top 31 probes subset that have been selected by the proposed *DV1* and *DV2* methods "intersection subset". The accuracy values obtained by Naïve Bayes, Random Forest, Hoeffiding Tree, SVM and Simple Logistic classifier using this subset are: 99.13%, 97.98%, 99.13%, 96.83% and 96.55%, respectively. These results show that cancer classification achieves lower predication accuracy than *DV1* or *DV2* or both due to missing information in intersection subset, and thus we confirm that the DNA methylation has several patterns that play significant role in human cancer. There is no single probes subset to identify these patterns and each feature selection method can provide different probes subset.

V. Conclusion and Future Work

Feature selection and extraction are of vital importance for accurate cancer classification, by skipping unnecessary information that introduce noises and decrease predication accuracy. This article proposes a framework based on novel feature selection methods along with extraction methods, to identify the informative probes that underlie the pathogenesis of tumor cell proliferation and improve cancer classification accuracy. The proposed feature selection method *DV1* uses statistical variation in terms of the standard deviation for obtaining the discriminative value while the other proposed feature selection method *DV2* uses entropy to rank features and hence obtains the more variational features with lower amount of uncertainty involved in its values. First, our framework uses *DV1* to identify the good marker probes subset, afterwards, in order to predict cancer, the average methylation density of three regions: hypomethylation, midmethylation and hypermethylation is calculated from the selected methylated probes as new features. The effectiveness of the proposed framework is evaluated by the breast cancer classification accuracy in probe regions, where the results are evidence that, our proposed framework has the ability to predict cancer using only three features out of 485577 features. As an example, SVM classifier gives 98.85% as higher prediction accuracy, and this highlights the importance of using feature selection and extraction methods in cancer classification issues based on DNA methylation.

Furthermore, observing probes subsets that have been selected from different feature selection methods confirmed that DNA methylation has several patterns and there is no single probes subset to identify these patterns. The results highlight the difference in methylation's behavior between the normal and abnormal samples in probes regions, and this difference confirms that the role of DNA methylation in cancer is to activate or silence some genes by decreasing or increasing their methylation respectively.

A new future work is identified based on the current study. We plan to improve the formula of feature extraction method instead of the current formula "average methylation density", to obtain higher results and improve cancer classification accuracy.

References

[1] Z. Herceg and P. Hainaut, "Genetic and epigenetic alterations as biomarkers for cancer detection, diagnosis and prognosis," Mol Oncol, vol. 1, pp. 26–41, 2007.

[2] M. Dawson and T. Kouzarides, "Cancer epigenetics: From mechanism to therapy," Cell, vol. 150, Issue 1, pp. 12–27, 2012.

[3] X. Ma and X. Gao, “Epigenetic Modifications and Carcinogenesis of Human Endometrial Cancer,” Austin Journal of Clinical Pathology, vol. 3, pp. 1-9, 2014.

[4] T. Mikeska and J. Craig, “DNA Methylation Biomarkers: Cancer and Beyond,” Genes, vol. 5, pp. 821-864, Sep 2014.

[5] Y. Ma, X. Wang and H. Jin, ”Methylated DNA and microRNA in body fluids as biomarkers for cancer detection,” Int J Mol Sci, vol. 14, pp. 10307–31, 2013.

[6] M. Pouliot, Y. Labrie, C. Diorio and F. Durocher,” The Role of Methylation in Breast Cancer Susceptibility and Treatment,” Anticancer Res, pp. 4569-74, September 2015.

[7] K. Williams, J. Christensen, and K. Helin, ”DNA methylation: TET proteins – guardians of CpG Islands?,” EMBO Rep, 13:pp. 28-35, 2012.

[8] I. Valavanis, E. Pilalis, P. Georgiadis, S. Kyrtopoulos and A. Chatziioannou, ”Cancer Biomarkers from Genome-Scale DNA Methylation: Comparison of Evolutionary and Semantic Analysis Methods,” Microarrays (2076-3905), vol. 4 , Issue 4, p647, December 2015.

[9] J. Li, H. Su and H. Chen, “Optimal Search Based Gene Selection for Cancer Prognosis,” in Proc. 11th Americas Conference on Information Systems, 2005.

[10] S.T. Li, C. Liao, and J.T. James, “Gene Feature Extraction Using T-Test Statistics and Kernel Partial Least Squares,” Springer, vol. 4234, pp. 11-20, 2006.

[11] B. Zheng, S. W. Yoon, and S. S. Lam, “Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms,” Expert Systems with Applications, vol. 41, no. 4, pp. 1476–1482, 2014.

[12] I. Kopriva, A. Jukić, and A. Cichocki, "Feature extraction for cancer prediction by tensor decomposition of 1D protein expression levels," in Proc. IASTED Conference on Computational Bioscience CompBio, pp. 277-283, July 2011.

[13] Y. Liu, U. Aickelin, J. Feyereisl, and L.G. Durrant, “Biomarker CD46 Detection in Colorectal Cancer Data based on Wavelet Feature Extraction and Genetic Algorithm,” Knowledge Based Systems, 2012.

[14] C. M. Fontes, R. Natowicz, R. Rouzier, and A. P. Braga, “Isolation of DNA Probes and Their Correlation with Neoadjuvant Chemotherapy outcome for Breast Cancer,” In: Congresso Brasileiro de Automática, 2011.

[15] D.L. Tong, “Genetic Algorithm-Neural Network: Feature Extraction for Bioinformatics Data,” Bournemouth University Research Online, 2010.

[16] J. Zhuang, M. Widschwendter, and A. Teschendorff, ”A comparison of feature selection and classification methods in DNA methylation studies using the Illumina Infinium platform,” BMC Bioinformatics ,Apr 2012.

[17] K. Anuradha, and K. Sankaranarayanan, “Comparison of Feature Extraction Techniques to classify Oral Cancers using Image Processing,” International Journal of Application or Innovation in Engineering, 2013.

[18] Z. Cai, D. Xu, Q. Zhang, J. Zhang, et al, ”Classification of lung cancer using ensemble-based feature selection and machine learning methods,” Mol. BioSyst, vol. 11 , pp. 791 -800, 2015.

[19] S. Sebastian, P. Justyna, and F. Krzysztof, ”Multiclass Classification Problem of Large-Scale Biomedical Meta-Data, “Procedia Technology, vol. 22, pp. 938-945, 2016.

[20] B. Baur and S. Bozdag, ”A Feature Selection Algorithm to Compute Gene Centric Methylation from Probe Level Methylation Data,” PLOS ONE, Edited by Jianhua Ruan, vol. 11, issue 2, Feb 2016.

[21] http://rnbeads.mpi-inf.mpg.de/

[22] V. Canedo, N. Maroño and A. Betanzos, ” Feature Selection for High-Dimensional Data,” Springer,2015.

[23] T. Cover, J. Thomas, ”Elements of Information Theory 2nd Edition,” Wiley Series in Telecommunications and Signal Processing, September 2006.

[24] S.J. Sheather, “Density Estimation,” Statistical Science, vol. 19, no. 4, pp. 588-597, 2004.