IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 19, 2018, accepted September 10, 2018, date of publication September 19, 2018, date of current version October 17, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2871027

Vertical and Horizontal DNA Differential Methylation Analysis for Predicting Breast Cancer

ABDULMAJID F. AL-JUNIAD[1,2], TALAL S. QAID[1,3], MOHAMMAD YAHYA H. AL-SHAMRI[1,2], (Member, IEEE), MAHDI H. A. AHMED[1,4], AND ABEER A. RAWEH[1,3]

[1]College of Computer Science, King Khalid University, Abha 61413, Saudi Arabia
[2]Faculty of Engineering and Architecture, Ibb University, Ibb, Yemen
[3]Faculty of Computer Science, Hodeidah University, Al Hudaydah, Yemen
[4]College of Engineering, Taiz University, Taiz, Yemen

Corresponding author: Mohammad Yahya H. Al-Shamri (mohamad.alshamri@gmail.com)

This work was supported by King Khalid University, Abha, Saudi Arabia, under Grant 71/2017.

ABSTRACT DNA methylation plays an important role for initiation and development of human cancers; therefore, it is used as a biological marker for early detection of cancer. A huge number of features for each sample and a low number of the available samples are two main problems of this field. This paper presents novel vertical, horizontal, and cascaded DNA methylation feature analysis methods in promoter regions. Vertical analysis processes each feature across all normal or cancer samples to get indicators about the methylation level. The generated values are used to select a subset of features within a given threshold. This set undergoes a horizontal analysis process where we group many features into a window that is used to yield a single value. Hence, the original sample size goes through two reduction steps: the first one is an unsupervised feature selection via the vertical analysis of the features and the second one is a feature extraction via the horizontal process of the selected features. For evaluation and comparison, we used traditional feature selection methods and SVD to compare them with the proposed approaches and found that the proposed approaches outperform all other approaches with a good margin. The results of vertical analysis or horizontal analysis alone are better than traditional approaches. Moreover, the results are improved more when combining both types of analysis. With only 97 features, the proposed combined approach is 99.16% accurate while the best traditional classification is only 98.16% accurate with 31 195 features. The combined approach achieved 8.8% to 54.3% improvement percentages compared to all other approaches in terms of a mean absolute error and a root-mean-square error. This indicates that the cascaded approach is far better than the previous approaches. Moreover, the combined approach improves the system accuracy and reduces space and processing complexities of the system.

INDEX TERMS DNA methylation, feature selection, feature extraction, cancer prediction, differential methylation.

I. INTRODUCTION

Cancer is responsible for 13% of global deaths and can develop anywhere in the body but all cancers are characterized by multiple genetic and epigenetic genomic alterations which lead to uncontrolled cell growth and reduced cellular differentiation [1]–[3]. In fact, epigenetics is the science that mainly studies external and environmental factors which activate or inhibit the work of genes and affect how the cell reads genes. The DNA methylation is an important epigenetic factor which plays an important role for the initiation and progression of human cancers and therefore could potentially be employed as a biomarker for early detection of cancer and as a predictor of treatment response. More specifically, the aberrant methylation of CpG islands in the promoter region is widely recognized as a tumor suppressor silencing mechanism in cancer [4], [5]. DNA methylation can change the DNA phenotype not its genotype where it can change the gene expression in cells when they divide from stem cells into a particular tissue cells. The changed gene expression stabled and the cell does not revert back to the stem cell or another type of cell [3]. Hence, understanding the molecular mechanisms of epigenetic alterations at the early stages of tumorigenesis may be very important in developing new cancer treatments. In fact, DNA methylation can be influenced by several factors including age, environment / lifestyle, and disease state [2], [3].

Computationally, the dataset of DNA methylation suffers from two significant problems, namely the high-dimensionality and high noises which are considered big challenges for classification. In addition, DNA methylation dataset has large number of features (genes or probes) with small number of samples. So, during training process this may lead to performance degradation of classification and raise the risk of overfitting [6].

One possible solution is to use feature selection to obtain the most relevant feature set and eliminate the redundant and irrelevant features [7], [8]. This set of discriminative features will play an important role for classifying normal samples from cancer ones. Sometimes, generating new features from the existing ones via feature extraction will be more efficient and sometimes combining feature selection with feature extraction will produce a more powerful method for classification [9]. This will reduce overfitting, improve prediction accuracy, decrease the time and space complexities of the classification process and open new directions for this field.

However, this is not an easy job for DNA methylation. The question that may arise 'Does the DNA methylation value at the feature level give useful information for classification?' The answer is "yes" but up to a very limited level. The methylation values are actually continuous and are different from sample to sample within the same sample set. Hence, the usefulness of these values as isolated values is very limited. A more logical approach is to accumulate these values at the feature level across all samples or at the sample level across many features to achieve more useful indicators.

This paper introduces both methods of dealing with the methylation values. It starts by accumulating the values at the feature level across all samples and hence finding the average value for each sample set alone. This process is done for both normal and cancer sample sets. The absolute difference between the two means is used as an indicator for feature selection or deselection. In the second step, the differences between some of the extracted features are accumulated horizontally within a specified window to produce a new feature. This extraction process uses the sum of differences between successive features within the window. Therefore, the differential analysis is utilized vertically to select some features and then horizontally to extract new features. Beside their better performance, our approaches minimize the processing time and the allocated memory while maintaining high accuracy.

The main contributions of this paper are:

- Building a framework for differential DNA methylation classification.
- Proposing a novel Differential Mean Feature Selection (DMFS) by utilizing the vertical analysis of features across the sample sets.
- Proposing a novel Differential Windowed Feature Extraction (DWFE) by utilizing the horizontal differential DNA methylation analysis which generates new compact and representative feature set.
- Proposing a cascaded DMFS-DWFE approach of vertical and horizontal analysis of DNA methylation.

The rest of this paper is organized as follows: the related work of DNA methylation and differential DNA methylation is given in Section II. The dataset and the examined methodologies are presented in Section III. In section 4, we analyze the obtained results and discuss them. The last section lists conclusions and highlights the possible future work directions.

II. RELATED WORK

The DNA methylation refers to the addition of a methyl (CH3) group to the cytosine or adenine nucleotides. This methyl group may be added to the fifth carbon atom of the cytosine base or the sixth nitrogen atom of the adenine base in the context of 5'-CG-3' (CpG dinucleotide) across human genome by DNA methyltransferase (DNMT) enzymes [3], [10]. Many studies linked the aberrant DNA methylation to cancer. However, most of these studies have been limited to the analysis of promoters and CpG islands (CGIs). Recently, new technologies for whole-genome DNAm (methylome) analysis have been developed [11], [12].

DNA methylation can be classified into hyper or hypo methylation. Many studies have associated hyper-methylation of tumor suppressor genes and hypo-methylation of oncogenes to the tumorigenic process [1]–[5], [13]. Hence, finding the hyper-methylated regions helps us to early discover the cancer [5], [14]. Ehrlich and Jiang [15] indicated that DNA hypo-methylation associated with cancer is probably as frequent as cancer-linked DNA hyper-methylation. They gave a caution to be used in development of treatment schemes for cancer involving DNA demethylation because they might result in increased tumor progression. Sproul *et al.* [16] suggested that the hyper-methylated gene does not directly contribute to cancer development via silencing. Instead aberrant hyper-methylation reflects developmental history and the perturbation of epigenetic mechanisms maintaining these repressed promoters in a hypo-methylated state in normal cells [17].

Ehrlich and Lacey [18] found that much more cancer-linked hypo-methylation of unique gene sequences and hyper-methylation of repeated sequences than previously found, although there are differences in the frequency with which subsets of sequences undergo hypo- or hyper-methylation. Tan *et al.* [10] indicated that aberrant DNA methylation is a frequent epigenetic event in pancreatic cancer. They identified 23 and 35 candidate genes that are regulated by hyper-methylation and hypo-methylation in pancreatic cancer, respectively. They also identified candidate methylation markers that alter the expression of genes critical to gemcitabine susceptibility in pancreatic cancer. Fukushige and Horii [19] described the mechanism that established and maintained DNA methylation patterns as well as the mechanism of aberrant gene silencing in cancer. Moreover, they introduced methods to isolate the DNA methylation biomarkers. They indicated that probably some combination of the various biomarkers, including genetic, epigenetic and serum ones will facilitate more reliable diagnosis, and DNA methylation biomarkers will be central to this development.

Model *et al.* [20] demonstrated how phenotypic classes can be predicted by combining feature selection and discriminant analysis. They showed that the right dimension reduction strategy is of crucial importance for the classification performance. Feltus *et al.* [21] and Previti *et al.* [22] used classifications algorithms to classify CpG islands. Zhuang *et al.* [23] highlighted the importance of tailoring the feature selection and classification methodology to the sample size and biological context of the DNA methylation study. Das *et al.* [24] described a computational pattern recognition method for both CpG islands and non-CpG island regions that is used to predict the methylation landscape of human brain DNA. A feature selection algorithm based on sequential forward selection was developed by Baur and Bozdag [25]. Their algorithm utilized different classification methods to compute gene centric DNA methylation using probe level DNA methylation data.

Zhou *et al.* [26] tried to find efficient feature selection methods to select a small number of informative genes using mutual information and rough sets. Nayyeri and Noghabi [27] proposed a sparse compact incremental learning machine for cancer classification on microarray gene expression data that is robust against diverse noises and outliers. Moghadam *et al.* [28] proposed a rule-based classifier to report combinations of CpG sites for identifying particular methylation changes in these sites. Kurdyukov and Bullock [29] gave an assessment of DNA methylation within particular regulatory regions/genes of interest. Wong *et al.* [7] presented a feature set reduction to enable a scalable feature selection on datasets with high dimensional data. They argued that this approach handles efficiently high resolution datasets that achieve better disease subtype classification of samples for potentially more accurate diagnosis and prognosis. This allows clinicians to make more informed decisions in regards to patient treatment options.

Hira and Gillies [8] tried to utilize prior knowledge to segment microarray datasets to identify candidate sets of genes for hypothesis testing. They divided the methylation dataset into subsets that contains only the probes that relate to a known gene pathway which will be used later independently for classification. Hira *et al.* [30] studied the relationship between response to the treatment and the features extracted from the measured methylation profiles to predict the outcome of a putative treatment regime. Ding and Peng [31] proposed a feature selection framework called a minimum redundancy – maximum relevance feature selection to provide a more balanced coverage of the space and capture broader characteristics of phenotypes.

List *et al.* [32] argued that a largely improved classification model can be obtained by combining methylation and gene expression data which reflects differences not only on the transcriptomic, but also on an epigenetic level. Jain *et al.* [33] used methylation profiles to classify four different kidney and lung cancer types with an accuracy exceeding 90% with only 16 features. Saeys *et al.* [34] gave a basic taxonomy of feature selection techniques, and discussed their use for bioinformatics applications. Li *et al.* [35] gave an optimal search-based subset selection method for high-dimensional gene array data that evaluate the group performance of genes and help to pinpoint global optimal set of marker genes. Li and Yin [36] proposed a multi-objective biogeography based optimization method to select the small subset of informative gene relevant to the classification.

Recently, Raweh *et al.* [37] utilized feature selection and proposed feature extraction methods for predicting cancer. They tried to analyze the set of selected features for further investigation where they found that the DNA methylation density of this set can clearly differentiate the cancer tissue from normal one.

III. MATERIALS AND METHODS

The Cancer Genome Atlas (TCGA) dataset from Max Planck Institute for Informatics (MPI) is used in this study [38]. The dataset contains several types of cancer: blood, breast, intestinal, brain and other types of cancer. For a specific gene, promotors are responsible for establishing transcription and are located near the transcription start sites of genes. There are 100 to 1000 base pairs for each promotor and many binding sites for the RNA polymerase complex may be there along its length. The degree of DNA methylation that extracted from the gene promoter regions are 31195 promoters. In this paper the breast cancer is examined where the number of samples are 598 (98 are normal samples and 500 are cancer samples), so we have a matrix of DNA methylation values consists of 598 rows and 31195 columns (features) in addition to the last column which is the class type with binary value (0 for normal sample or 1 for cancer sample).

Figure 1 shows the DNA methylation density for breast normal and cancer tissues. The density values show three different methylation levels, low, medium, and high methylation levels. The low methylation values correspond to hypomethylation that activates genes while high methylation values correspond to hypermethylation that makes genes silent [37]. The density graph shows that hyper and hypo methylated features are more than medium methylated features for both normal and cancer tissues. This drives the classification to good accuracy results. For further improvements, we need to capture the minor differences between methylation values of normal and cancer ones.

In general, the two density graphs look similar except their thickness. This indicates that the individual methylation values are very close and usually the difference between

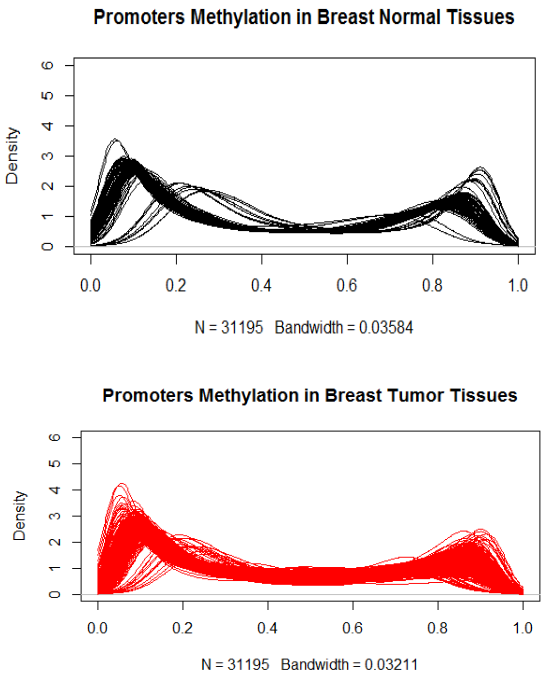

FIGURE 1. DNA methylation density for both normal and cancer tissues of breast [37].

the normal and cancer values is very small which cannot be captured easily. The DNA methylation values for cancer tissues occupy wider range than that of normal one but with a very small margin. This makes it difficult for many machine learning techniques to differentiate between them for further improvement.

A. SIMPLE CLASSIFICATION

There are many classification techniques and each one of them has its own pros and cons. This paper will test three different classification techniques, namely; Naïve Bayes classifier, random forest classifier, and support vector machine classifier. The Naïve Bayes classifier is the direct and simplest classifier that utilizes Bayes theorem for conditional probabilities of random variables given known observations to build the classifiers. In this classifier all features are assumed independent from each other and it calculates independently the probability of each feature for a particular class label [39], [40]. This classifier is simple and computationally fast to reach a decision. The disadvantage of this classifier is that it assumes a specific form for the feature probability distribution of each class.

The second examined classifier is Random forest which is an ensemble predictor close to the nearest neighbor predictor. The ensemble predictors assume that strong predictors can come up from weak ones. Starting with decision trees with controlled variance as weak predictors, random forest goes ahead and combines these weak predictors to form an ensemble. The advantages of this classifier are robustness, no requirement for normalization, and immunity to collinearity [41].

The third classifier is the support vector machine which is a supervised learning process. It uses a non-linear mapping to map the input vectors into some high dimensional feature space Z. The transformed feature space of SVM classifier needs a kernel function to fit a maximum-margin hyperplane in it. This transformation is usually high dimensional and uses nonlinear polynomial or radial basis kernel functions [42]. The SVM classifier is computationally expensive but with high prediction accuracy when compared to other classifiers [39]. Two parameters are associated with the SVM training, namely, the cost and the kernel function parameters. For our experiments we set the batch size to 100, and the cost to 1. In terms of kernel functions, we test both polynomial and radial basis kernel functions and use the one giving the best results.

For our experiments, R language (RStudio 1.1.453) and WEKA 3.8.1 tools are used as implementation medium. Two metrics are used for evaluating the prediction quality of our experiments, namely, mean absolute error (MAE), and root mean square error (RMSE). The mean absolute error is given by [43]:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|p_i - a_i| \tag{1}$$

where p_i is the predicted class, a_i is the actual class and n is the number of tested samples. The root mean square error is defined as [43]:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - a_i)^2} \tag{2}$$

Moreover, we use two metrics for measuring the model performance namely, classification accuracy and F-measure. Accuracy is the ratio of the number of correct predicted classes to the total number of tested samples. Both measures will be calculated directly from the confusion matrix which provides us with a complete view about the model performance [44], [45]:

$$\text{Accuracy} = \frac{t_p + t_n}{t_p + f_p + t_n + f_n} \tag{3}$$

Here, t_p and t_n are the numbers of true positive and true negative cases respectively. Similarly, f_p and f_n are the numbers of false positive and false negative cases respectively. Accuracy works well if we have equal number of samples for each class and the cost of misclassification of cancer is very high. Therefore, we will use also F-measure which will tell how precise and robust the classifier is. The greater the F measure, the better is the model performance.

F-measure is a good measure to get a balance between precision and recall for uneven class distribution as for Breast cancer where we have 98 normal samples and 500 cancer samples. A balanced F-measure is a single score due to a harmonic mean of precision and recall which can be any value

TABLE 1. Results of simple classification approach.

Classification Method	Accuracy	F-Measure	MAE	RMSE
Naive Bayes Classifier	96.4883	96.6	0.0351	0.1874
Random Forest	97.9933	98.0	0.0629	**0.132**
SVM	**98.1605**	**98.2**	**0.0184**	0.1356

between 0 and 1 [44], [45].

$$F = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2t_p}{2t_p + f_p + f_n} \tag{4}$$

The results of the above mentioned classifiers for breast tissues are depicted in Table 1.

The results show that SVM is the best in terms of accuracy, F-measure and MAE while random forest is the best in terms of RMSE. This indicates that the system can identify the cancer from normal cases with 98.16% with only 0.0184 as MAE. Hereafter, we will use SVM for the classification purpose whenever it is required because it is the best among the examined classification ones.

B. FEATURE SELECTION

It is beyond doubt that the feature selection is very important for large scale data but unfortunately the best method does not exist. The researchers try either to find a good method for a specific problem setting or to merge many methods for a hybrid approach [46]. The main goal for using feature selection techniques is to reduce the number of attributes in the dataset by including only useful features in the dataset without changing them. This is usually done by selecting relevant features that may describe properly the problem on hand and discarding irrelevant ones without affecting the system performance which has to be within an acceptable range [47], [48].

We used three algorithm-independent feature selection methods, namely Information gain, Correlation-based, and ReliefF. These methods are simple, fast, and able to handle large-scale datasets [49]. The Information Gain filter considers a single feature at a time and evaluates the features according to their information gain. Whereas the Information Gain filter is univariate method, the Correlation-based is multivariate filter algorithm that uses a correlation-based heuristic evaluation function to rank feature subsets. The ReliefF filter which is an extension to the original Relief algorithm based on randomly selecting an instance from the data and then locating its nearest neighbor from the same and opposite class [49].

Feature selection methods return the features as an ordered list. The first feature in this list has the highest weight and represents the highest importance to describe the problem in hand. Usually, selecting a representative set of features from the generated ordered list requires manual trial and error search which may give different results for various applications. For our experiments, we got it 9% for some methods and 3% for others therefore we test all methods with three selection percentage values 3%, 6% and 9%. Low selection percentages show that the employed method can identify the discriminative features efficiently.

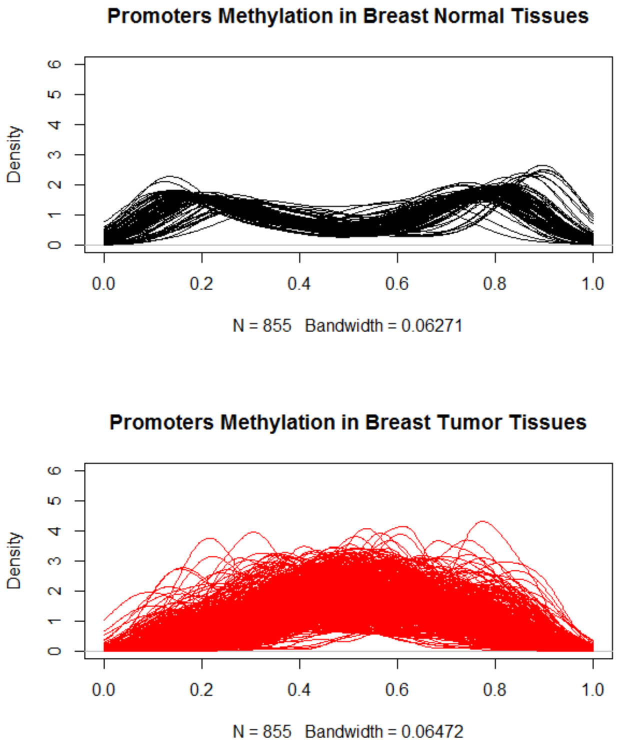

FIGURE 2. DNA methylation density for both normal and cancer tissues of breast due to feature selection.

The DNA methylation density graph, after selecting some discriminative features, for both normal and cancer tissues are given in Figure 2. The density graph for normal samples follows to a large extent the same way of original density graph of Figure 1. The difference depends only on the methylation level for the hyper and hypo methylated features which now become smaller. On the other hand, the density graph for cancer samples differs totally from that of original cancer samples. The medium methylation values now dominate the density graph and become very important for the classification process.

Numerically, the results of the examined feature selection methods in Table 2 show that Information Gain and ReliefF methods give good results with only 9% of the features, which are around 2808 out of 31195. These results are better than those of SVM simple classification by 0.6814%, 0.611%, 36.41% and 20.2% in terms of accuracy, F-measure, MAE, and RMSE respectively. We follow the same formulae given in [43] for calculating improvement percentages. This indicates that many features are redundant and do not contribute to the classification process if not misguide the classification process and behave like outliers. Another important point here is that the system complexity is reduced a lot and is now about 9% only of its original one.

C. SINGULAR VALUE DECOMPOSITION ANALYSIS

One may argue that the DNA methylation microarray may not lead to good results due to the noise in data and the

TABLE 2. Results of feature selection approach.

		Information Gain	Correlation -based	ReliefF
Accuracy	3%	98.6622	98.6622	98.6622
	6%	98.6622	98.6622	98.6622
	9%	**98.8294**	98.6622	**98.8294**
F-Measure	3%	98.7	98.7	98.7
	6%	98.7	98.7	98.7
	9%	**98.8**	98.7	**98.8**
MAE	3%	0.0134	0.0134	0.0134
	6%	0.0134	0.0134	0.0134
	9%	**0.0117**	0.0134	**0.0117**
RMSE	3%	0.1157	0.1157	0.1157
	6%	0.1157	0.1157	0.1157
	9%	**0.1082**	0.1157	**0.1082**

limited range of the values. Hence we must find new ways to change the matrix internal structure. In general, there are two approaches to change the internal structure of the DNA methylation microarray, namely feature extraction and matrix decomposition. In this section, we will employ singular value decomposition (SVD) which is actually Eigen decomposition. SVD is useful for high-dimensional matrices because it can be used for dimensionality reduction. The result of this procedure is a three low-dimensional matrices, namely, left-singular (I), singular (D) and right-singular (W) matrices respectively [50], [51]. The process of decomposing the original matrix to its basic components will be useful for identifying the most important features for that matrix. The eigenvalue vector of the decomposed matrix is the most important basic component of the decomposition process which can be employed to generate a new version of that matrix. The decomposition process sorts the eigenvalues for the matrix which can be as many as the matrix size or only a little number.

Systematically, we decompose the microarray matrix A_{QN} (the class column is not included, only features are included in microarray matrix A_{QN}) using SVD into three matrices I_{QP}, D_{PP} and W_{PN}. This paper uses the first two matrices for generating a reduced matrix by selecting only some useful eigenvalues. Hence, the dimension P of the two matrices (D_{PP} and I_{QP}) is reduced into L by selecting only L eigenvalues from the matrix D_{PP} and the corresponding columns in matrix I_{QP}. After that, the reduced matrix $\boldsymbol{A_{\bar{Q}L}}$ is reconstructed by matrix-matrix multiplication as follows:

$$\boldsymbol{A_{\bar{Q}L} = I_{QL}{}^{*}D_{LL}} \tag{5}$$

Next, the class column is added to the reduced matrix $\boldsymbol{A_{\bar{Q}L}}$ to get a new matrix which is used as an input to the classifier. Decomposing the DNA methylation microarray using SVD and thereafter selecting a predefined percentage of its eigenvalues will allow us to see the effect of changing the internal structure of the matrix for the classification process. The new values will highlight the importance of the new features for predicting the cancer tissues.

TABLE 3. Results of SVD approach.

Percentage	Accuracy	F-Measure	MAE	RMSE
5%	**98.662**	**98.7**	**0.0134**	**0.1157**
10%	98.495	98.5	0.0151	0.1227
20%	98.328	98.4	0.0167	0.1293
30%	98.161	98.2	0.0184	0.1356

For our experiments, we use the top 5%, 10%, 20%, and 30% of the eigenvalues (the total number of eigenvalues after applying SVD is r = 598) for generating the new matrix. As listed in Table 3, the results show that the top 5% eigenvalues give the best results among others. For comparison, the results are better than SVM classification but less than that of feature selection experiments by very small margins. However, the number of features here is very small, i.e. only 30. This will give very good time and space complexities for such approach with almost the same prediction results as feature selection.

D. VERTICAL AND HORIZONTAL DNA METHYLATION ANALYSIS

Apparent promoter methylation plays a critical role in human breast carcinogenesis because it occurs at the early stage of breast tumor. High methylation value may silence tumor suppressor genes which lead to cell growth and hence to the genesis of neoplasia like breast tumorigenesis. Therefore it is used as a potential marker for early diagnosis and therapeutic of breast cancer. However, methyaltion values are usually scaled between 0 and 1 and no clear threshold can be obtained for identifying normal and cancer related values. The best way to deal with this complex problem is to relate values to each other and see if there is any difference. Hence, identifying those features having remarkable differences will be a very good input to the classification process. The difference in the methylation values can be considered as an important biomarker for determining cancer. This will lead to differential methylation approaches which can be processed vertically among cancer and normal samples of the same tissue or horizontally among different features of the same sample.

In fact, most of the previous work deals with methylation values separately while others try to find the differential methylation regions biologically and then make a computational analysis on those regions. However, we argue that the differences between DNA methylation values will be more useful, efficient and will give better indicators. The main idea here is to identify differential methylation values vertically and horizontally to maximize the utilization of this biomarker for predicting breast cancer.

The vertical DNA methylation analysis is done on all samples to select the most discriminative features which will discriminate cancer samples from normal samples as shown in Figure 3. Actually, vertical analysis tries to exploit the differences in methylation values between normal and cancer

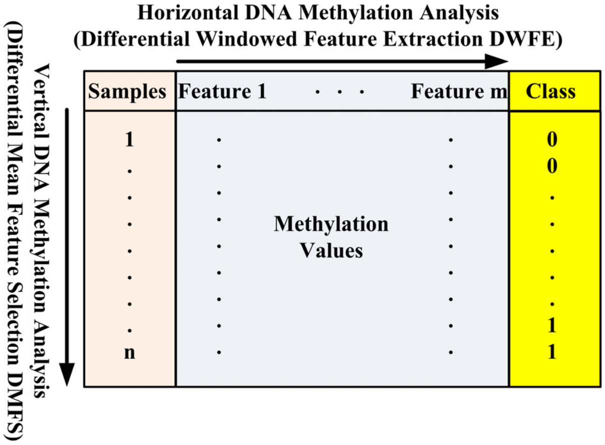
Horizontal DNA Methylation Analysis
(Differential Windowed Feature Extraction DWFE)
Vertical DNA Methylation Analysis
(Differential Mean Feature Selection DMFS)
Samples
Feature 1
Feature m
Class
1
n
Methylation Values
0
0
1
1

FIGURE 3. Idea of vertical and horizontal DNA methylation analysis.

samples. On the other hand, the horizontal DNA methylation analysis tries to extract new features based on differences of methylation values across each sample. The sum of absolute differences within a window will be the new feature and will be a good indicator across that window. The following subsections discuss the proposed vertical and horizontal analysis separately and highlight the strengths of each one of them for predicting cancer tissues. The last subsection discusses the proposed combined approach.

E. DIFFERENTIAL MEAN FEATURE SELECTION (DMFS)

We analyze vertically each feature across all samples and find the mean value of the DNA methylation values of both normal and cancer samples separately. These feature-wise means represent global descriptors for all features over normal or cancer samples. Now a rough estimation of the usefulness of each feature for classifying samples into normal or cancer ones can be developed using the absolute difference between the mean values of normal and cancer samples. So we will call this novel approach as Differential Mean Feature Selection (DMFS). The means for both cancer and normal tissues will reflect the minor differences between the two samples at the feature level. It is true that the mean may show some compensation behavior but this will be for usual features. For unusual features, the DNA methylation will be low or high and hence it will have a clear impact on the mean value.

A threshold value to select the feature or not is used for this unsupervised feature selection process. The threshold value indicates that the difference between the two mean values of normal and cancer samples is adequate to say that feature is discriminative for cancer classification. However, the main important point here is how to identify the appropriate threshold value? For our work we use trial and error policy to estimate this value. However, someone can utilize other approaches to do that. Our criterion for selecting this value is to test the system many times and selects the threshold value that gives the best results. Hence, this will depend on the search space and the cancer type. Figure 4 illustrates the process of DMFS where the two vectors containing the average of normal samples and the average of tumor samples are first calculated. After that the absolute difference vector is calculated based on the two mean vectors. The resulted vector is named weighting vector. Finally, the weighting vector is used to select features which correspond to a weight larger than a certain predefined threshold.

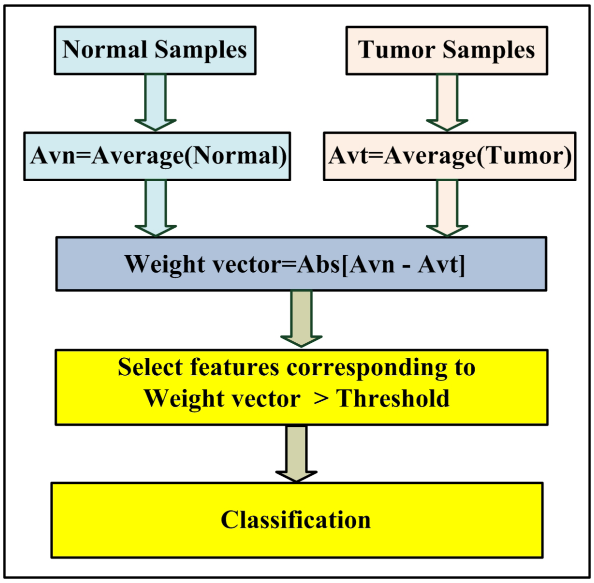
Normal Samples
Tumor Samples
Avn=Average(Normal)
Avt=Average(Tumor)
Weight vector=Abs[Avn - Avt]
Select features corresponding to Weight vector > Threshold
Classification

FIGURE 4. Vertical DNA methylation analysis (DMFS approach).

Mathematically, the set of features for each sample is:

$$F = \langle f_1, f_2, \ldots, f_m \rangle \tag{6}$$

The cardinality of this set is m. Assume a_i^n and a_i^c are means of feature i across normal and cancer samples, respectively. Accordingly, we can define the feature-wise vector means for normal and cancer samples as:

$$\begin{aligned} A^n &= \left\langle a_1^n, a_2^n, \ldots, a_m^n \right\rangle \\ A^c &= \left\langle a_1^c, a_2^c, \ldots, a_m^c \right\rangle \end{aligned} \tag{7}$$

Based on this we can define the differential methylation of feature i as:

$$dm_i = \left| a_i^n - a_i^c \right| \tag{8}$$

Alternatively, the differential methylation vector is:

$$WV = \left| A^n - A^c \right| = \left\langle dm_1, dm_2, \ldots, dm_m \right\rangle \tag{9}$$

WV is the weighting vector used for selecting features for the classification process. The system selects the features, with weights greater than a predefined threshold (THR), which will be used for the classification process. Usually, the threshold value is tuned to get the best prediction accuracy. Hence, the set of selected features will be:

$$FS = \left\{ f_{\sigma(i)} : dm_{\sigma(i)} \geq THR \right\} \tag{10}$$

TABLE 4. Results of DMFS approach.

Selection Percentage	Accuracy	F-Measure	MAE	RMSE
3%	**98.9967**	**99.0**	**0.0100**	**0.1002**
6%	98.8294	98.8	0.0117	0.1082
9%	98.8294	98.8	0.0117	0.1082

where σ is a permutation that orders features based on their differential methylation values such that $dm_{\sigma(1)} \geq dm_{\sigma(2)} \geq \ldots \geq dm_{\sigma(m)}$.

The results of DMFS approach are listed in Table 4. We can notice that it performs very well even with just 3% of the features, which are around 936 out of 31195. This indicates that 97% of the features are redundant and do not contribute to the classification process. Moreover, the system complexity is reduced by a very good factor and it is now only 3% of its original one.

F. DIFFERENTIAL WINDOWED FEATURE EXTRACTION (DWFE)

Feature extraction explores the features of the dataset under consideration and tries to extract hidden information and hence creates new features [52]. In literature, some authors treat feature selection and feature extraction interchangeably as the feature selection process which is the process of extracting relevant features from the available ones. However, in this paper we will differentiate between them and consider the process as a feature selection process if only some features are selected from the available ones according to some criteria and consider the process as a feature extraction process if some mathematical or logical operation is applied on available features to build new ones.

Actually, isolated methylation values may not give useful information out of thousands of such values. A more logical approach is to compare them with each other and see the difference between them. This will detect if there is any sudden increment or decrement in the values especially in the hyper or hypo-methylation regions. For this purpose, we set the window size to a predefined value and sum the absolute differences between successive values in the window and set the sum as a new extracted feature for that window. This way, we will gain the following benefits:

- The number of features is decreased by the window size factor.
- The differential movement of methylation values is captured within the window.

Figure 5 shows the process of horizontal DNA methylation analysis which is called Differential Windowed Feature Extraction (DWFE). In this approach, each sample is divided into a number of windows equal to:

$$d = \left\lceil \frac{m}{W} \right\rceil \tag{11}$$

where m is the number of features in each sample and W is a predefined window size.

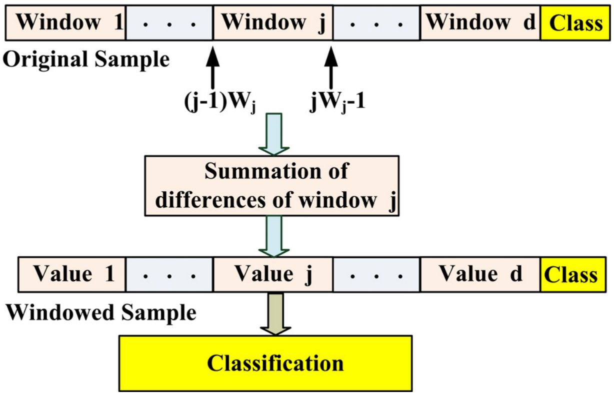

FIGURE 5. Horizontal DNA methylation analysis (DWFE approach).

For each window, W_j, the summation of differences is used to produce one value $V(j)$ for this window as follows:

$$\mathrm{V}(j) = \sum_{i=0}^{W_j-2} |f_{i+1} - f_i| \tag{12}$$

where $j = 1, \ldots, d$ is the window index, W_j is the actual size of window j, and $i = 0, \ldots, W_j - 1$ is the feature index within window j.

$$W_j = \begin{cases} W & j < d \\ m\%d & j = d \end{cases} \tag{13}$$

DWFE extracts d features from m DNA methylation features. The window size maybe tuned to get the best prediction accuracy.

The results of this approach for breast tissues are 98.8295%, 0.988, 0.0117, and 0.1082 in terms of accuracy, F-measure, MAE, and RMSE respectively. These results are similar to that of feature selection with only 312 extracted features that represent only 1% of the original sample size and only 11.11% of the size of feature selection approach. This is a very big achievement in reducing the cardinality of the feature set besides keeping the same system accuracy.

G. CASCADDED DMFS-DWFE

The results of the proposed DMFS and DWFE are promising and indicate their good ability to classify breast cancer tissues. This encouraged us to go one step further by exploring the combined effect of the vertical and horizontal differential methylation as a biomarker to predict breast cancer. In this direction, we propose a cascaded DMFS-DWFE approach, which works as follow: First, DMFS is applied on the whole dataset as mentioned before to select some discriminative features on which we will apply the next step. After that DWFE is applied on the set of the selected features resulted from the vertical analysis. The process needs two input values which are the selection threshold value and the window size. The values of the threshold and the window size are tuned to get the best prediction accuracy.

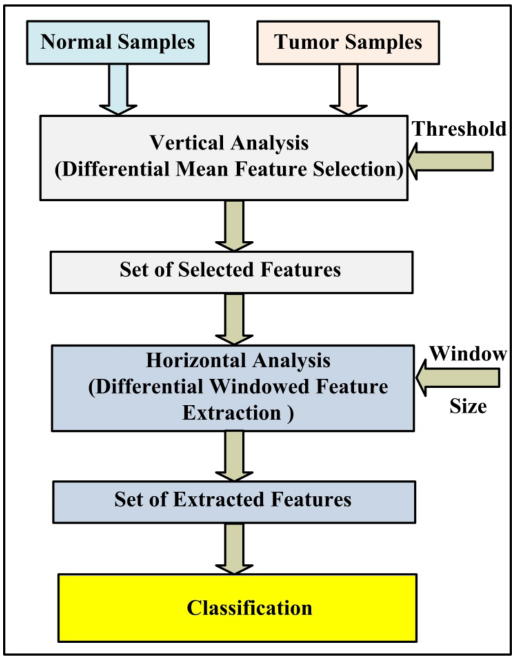

FIGURE 6. Cascaded DMFS-DWFE approach.

TABLE 5. Results of cascaded DMFS-DWFE approach.

Percentage	Accuracy	F-Measure	MAE	RMSE
10%	97.9933	98.0	0.0201	0.1417
20%	98.3278	98.3	0.0167	0.1293
30%	**99.1639**	**99.2**	**0.0084**	**0.0914**

Figure 6 shows the whole cascaded process where the proposed DMFS is applied on all features using different threshold values to select different percentages of features 10%, 20%, and 30% as listed in Table 5. We increased the selection percentages here compared to the previous approaches because we have another layer of refinement for features and therefore we have to keep enough number of features. For each percentage of the selected features, the proposed DWFE is applied with window size of 100. The results of the cascaded approach show that it is the best among all studied approaches. The following highlights some important points:

- This approach takes only tens of features which are very small compared to the original number of features.
- The results of this approach with only some tens of features are better than individual DMFS and DWFE approaches.

IV. RESULTS ANALYSIS AND DISCUSSIONS

As a classification method, SVM shows good results with respect to other examined classification methods. It achieves 98.2% for identifying cancer samples correctly. Further improvement is achieved by ReliefF feature selection method which achieves 98.8% for identifying cancer samples correctly with only 9% of the original features. In this paper, we go further by investigating the benefit of Eigenvalues for the classification process where we used SVD for decomposing the original matrix and utilizing the top 5% useful Eigenvalues for generating new matrix and then employing it for the classification process. This experiment achieves 98.7% accuracy in classifying cancer samples which is better than simple SVM classification but less than ReleifF results.

TABLE 6. List of experiments for all approaches.

Approach	Experiment(s)
Simple Classification	Naive Bayes Classifier
	Random Forest
	SVM
Feature Selection	Information Gain (3%, 6%, and 9%)
	Correlation-based (3%, 6%, and 9%)
	ReliefF (3%, 6%, and 9%)
SVD	Singular Value Decomposition (select 5%, 10%, 20% and 30% of the eigenvalues)
DMFS	Differential Mean Feature Selection (3%, 6%, and 9%)
DWFE	Differential Windowed Feature Extraction (Extract 312 features with window size = 100)
Cascaded DMFS-DWFE	Cascaded DMFS-DWFE (select 10%, 20%, and 30% features using DMFS then extract features from them with window size = 100).

The achieved results of the proposed approaches outperform the traditional approaches in overcoming the high-dimensionality of the DNA methylation data, which is usually higher than the number of collected samples. In addition to that, our approaches identify the most discriminative features for accurate cancer prediction. To approve that, six sets of experiments were conducted in this paper, one set for each approach. The experiment set of some approaches consists of many experiments based on the employed methods for that approach. There are some predefined parameters for some experiments that are discussed inside them. For summary, Table 6 depicts the considered approaches and lists the experiment(s) for each one.

The best results for all approaches are listed in Table 7 along with the number of features for each of them. The results show that applying the proposed approaches gives a good improvement. The improvement value depends on the method used and the applied approach. The results exhibit that DMFS outperforms simple SVM classification, SVD, ReleifF feature selection, and DWFE with only 3% of the features, i.e. 936 out of 31195 features. The improvements compared to simple ReliefF feature selection are 0.17%, 0.2%, 14.53%, and 7.4% in terms of accuracy, F-measure, MAE, and RMSE respectively. The improvements are very good with regards to errors which indicate that the system predictions about the class are very close to the actual ones. This is very important for cancer prediction as usually failing to predict the cancerous case will be very dangerous for the patient.

TABLE 7. Best results for all examined approaches.

Approach	Number of features	Features %	Accuracy	F-Measure	MAE	RMSE
SVM	31195	100%	98.1605	98.2	0.0184	0.1356
ReleifF	2808	9%	98.8294	98.8	0.0117	0.1082
SVD	30	0.096%	98.6620	98.7	0.0134	0.1157
DMFS	936	3%	98.9967	99.0	0.0100	0.1002
DWFE	312	1%	98.8294	98.8	0.0117	0.1082
Cascaded DMFS-DWFE	**97**	**0.31%**	**99.1639**	**99.2**	**0.0084**	**0.0914**

TABLE 8. Improvement percentages of cascaded DMFS-DWFE approach compared to other approaches.

Approach	Accuracy	F-Measure	MAE	RMSE
SVM	1.0222034	1.0183299	54.347826	32.595870
SVD	0.5087065	0.5065856	37.313433	21.002593
ReleifF	0.3384620	0.4048583	28.205128	15.526802
DWFE	0.3384620	0.4048583	28.205128	15.526802
DMFS	0.1688945	0.2020202	16	8.7824351

DMFS shows 98% accuracy to tag each sample to its correct class while it shows 99% to classify correctly cancer samples which demonstrates very high classification performance. These values jump to 99.16% accuracy and 99.2% F measure with the proposed combined approach which indicates that our model is very precise for predicting the sample class and the true positive cancer cases. This is very important here as we do not want to misclassify any caner case as normal case.

For more clarification, Table 8 illustrates the improvement percentages of the best approach (cascaded DMFS-DWFE) compared to all other approaches. The results show that there is a small improvement in accuracy compared to all other approaches which is around 1% only. The same thing is achieved for F-measure which has nearly the same improvement percentage. However in terms of MAE and RMSE there are very good improvements compared to other approaches. The range of these improvements is 8.8% to 54.3%. This indicates that the cascaded approach is far better than the previous approaches. The low value of prediction error indicates that prediction quality is very good and the system can predict the class of each sample with high accuracy.

Graphically, Figure 7 illustrates the improvement percentages of cascaded DMFS-DWFE approach compared to other approaches in terms of MAE and RMSE. The results show that differential methylation is very strong in identifying the cancer tissues. Sometimes, the difference is very small between the methylation values of successive features but if captured it will be very useful for classification purposes.

The required number of features for each approach is given in Table 7. This number is very important especially for DNA methylation where thousands of features are usually there. The results say that SVD is the best in terms of the utilized number of features where it is only 0.096% of the original size. However, its accuracy is somehow equal to that of simple feature selection. On the other hand, cascaded DMFS-DWFE has only 97 features where it is 0.31% of the original size but with superiority over the other approaches in all aspects.

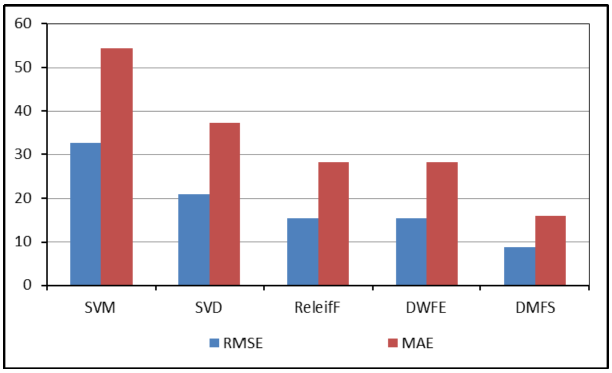

FIGURE 7. Improvement percentages of cascaded DMFS-DWFE approach compared to the other approaches in terms of MAE and RMSE.

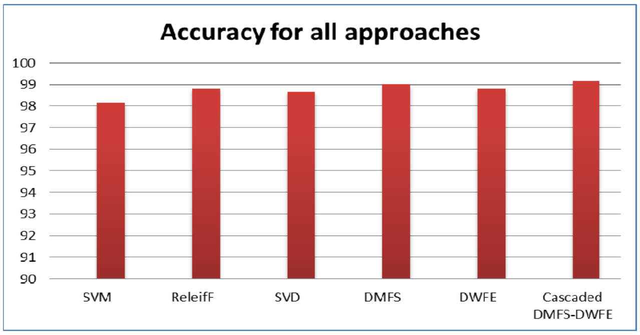

FIGURE 8. Accuracy comparison of all approaches.

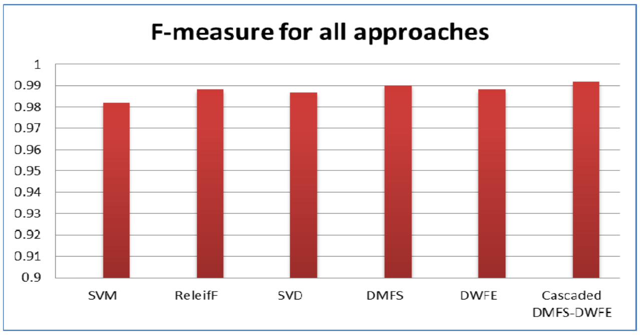

FIGURE 9. F-Measure comparison of all approaches.

In fact, the number of features for both approaches is very low compared to other approaches and hence we overcome the main difficulty of such applications which is the dimensionality size of the dataset. The second best one in this regard is DWFE which has only 312 features and can be adjusted more based on the nature of the dataset. Moreover, its accuracy is very good compared to many other approaches. This illustrates that our approaches are good in terms of accuracy, time and space complexities which make it appropriate for such high dimensional microarray data.

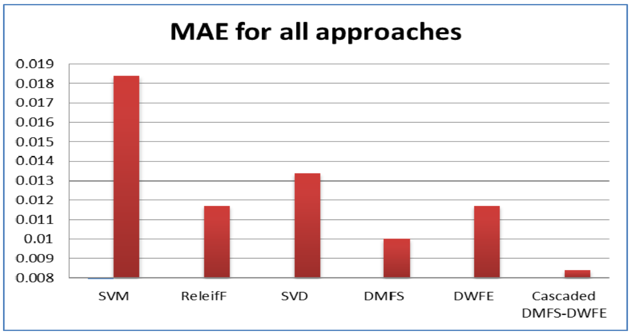

FIGURE 10. MAE comparison of all approaches.

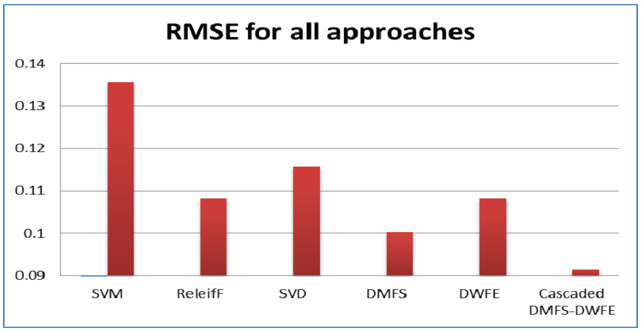

FIGURE 11. RMSE comparison of all approaches.

Figures 8 to 11 show the effect of differential DNA methylation analysis when we compare it to other approaches. The error is very small for the cascaded approach which indicates that the prediction values are either the same or very close to the original one. That means differential methylation is very effective for classifying tissues.

The prediction accuracy is improved with very less number of features and therefore the prediction time is reduced so much. In fact, the system complexity can be discussed in terms of space and time processing. Table 7 illustrates that the number of features is reduced so much with the proposed approaches which in turn will reduce the processing time so much since the system will use only these features at the classification stage. The offline stage for generating the set of features will be hidden from the user and can be done anytime or regularly if the samples are modified each time. Moreover, we need to store only that number of features for the classification purpose.

V. CONCLUSIONS

DNA methylation is an epigenetic indicator for activating or silencing a gene. The proposed framework exploits the DNA differential methylation vertically and horizontally to predict breast cancer more accurately. Vertically, DMFS utilizes differential mean between the mean of normal and cancer samples. The differential mean is used to select the most discriminative features above a certain threshold. Horizontally, the DWFE extracts new features based on the summation of absolute differences of DNA methylation values within a certain window. In general, the results of the proposed DMFS and DWFE approaches show improvements over simple classifier, traditional feature selection methods and SVD in terms of accuracy, F-measure, MAE, and RMSE.

In addition, the cascaded DMFS-DWFE approach exploits the combined effect of DMFS and DWFE approaches. The results of the cascaded DMFS-DWFE approach are the best among all examined approaches in this paper especially in terms of MAE and RMSE. The combined approach is very good also in terms of the time and space complexities as the number of features is reduced by a very significant factor. This is a very good achievement for such huge and high dimensional dataset. The dataset cardinality is reduced from 31195 to only 97 with the cascaded approach.

The proposed feature extraction can be modified and tuned in terms of the employed method for calculating the new feature or tuning the window size. Moreover, the feature selection percentages and window size can be calculated automatically based on some factors of the dataset like mean methylation value within the window which may be explored in future work.

REFERENCES

[1] A. Jones *et al.*, "Role of DNA methylation and epigenetic silencing of *HAND2* in endometrial cancer development," *PLOS Med.*, vol. 10, no. 11, p. e1001551, Nov. 2013.
[2] L. Moore, T. Le, and G. Fan, "DNA methylation and its basic function," *Neuropsychopharmacology*, vol. 38, pp. 23–38, Jul. 2013.
[3] C. Bock, "Epigenetic biomarker development," *Epigenomics*, vol. 1, no. 1, pp. 99–110, 2009.
[4] 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, no. 5, pp. 10307–10331, 2013.
[5] H.-C. Lai *et al.*, "DNA methylation as a biomarker for the detection of hidden carcinoma in endometrial atypical hyperplasia," *Gynecol. Oncol.*, vol. 135, no. 1, pp. 552–559, 2014.
[6] H. Hijazi and C. Chan, "A classification framework applied to cancer gene expression profiles," *J. Healthcare Eng.*, vol. 4, no. 2, pp. 255–283, 2013.
[7] G. Wong, C. Leckie, and A. Kowalczyk, "FSR: Feature set reduction for scalable and accurate multi-class cancer subtype classification based on copy number," *Bioinformatics*, vol. 28, no. 2, pp. 151–159, 2012.
[8] Z. M. Hira and D. F. Gillies, "Identifying significant features in cancer methylation data using gene pathway segmentation," *Cancer Inform.*, vol. 15, pp. 189–198, Sep. 2016.
[9] A. A. Raweh, M. Nassef, and A. Badr, "Feature selection and extraction framework for DNA methylation in cancer," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 7, pp. 30–36, 2017.
[10] A. Tan *et al.*, "Characterizing DNA methylation patterns in pancreatic cancer genome," *Mol. Oncol.*, vol. 3, nos. 5–6, pp. 425–438, 2009.
[11] A. Feber *et al.*, "Comparative methylome analysis of benign and malignant peripheral nerve sheath tumors," *Genome Res.*, vol. 21, no. 4, pp. 515–524, 2011.
[12] H. Alvarez *et al.*, "Widespread hypomethylation occurs early and synergizes with gene amplification during esophageal carcinogenesis," *PLoS Genet.*, vol. 7, no. 3, p. e1001356, 2011.
[13] T. Phillips, "The role of methylation in gene expression," *Nature Educ.*, vol. 1, no. 1, p. 116, 2008.
[14] C.-X. Song and C. He, "Balance of DNA methylation and demethylation in cancer development," *Genome Biol.*, vol. 13, p. 173, Oct. 2012.
[15] M. Ehrlich and G. Jiang, "DNA hypo-vs. hypermethylation in cancer: Tumor specificity, tumor progression, and therapeutic implications," in *DNA Methylation and Cancer Therapy*. Norwell, MA, USA: Kluwer, 2005, ch. 3, pp. 31–41.
[16] D. Sproul *et al.*, "Tissue of origin determines cancer-associated CpG Island promoter hypermethylation patterns," *Genome Biol.*, vol. 13, no. 10, p. R84, 2012.

[17] J. Sandoval and M. Esteller, "Cancer epigenomics: Beyond genomics," *Current Opinion Genet. Develop.*, vol. 22, no. 1, pp. 50–55, 2012.

[18] M. Ehrlich and M. Lacey, "DNA hypomethylation and hemimethylation in cancer," *Adv. Exp. Med. Biol.*, vol. 754, pp. 31–56, 2013.

[19] S. Fukushige and A. Horii, "DNA methylation in cancer: A gene silencing mechanism and the clinical potential of its biomarkers," *Tohoku J. Exp. Med.*, vol. 229, no. 3, pp. 173–185, 2013.

[20] F. Model, P. Adorján, A. Olek, and C. Piepenbrock, "Feature selection for DNA methylation based cancer classification," *Bioinformatics*, vol. 17, no. 1, pp. S157–S164, 2001.

[21] F. A. Feltus, E. K. Lee, J. F. Costello, C. Plass, and P. M. Vertino, "DNA motifs associated with aberrant CpG island methylation," *Genomics*, vol. 87, no. 5, pp. 572–579, 2006.

[22] C. Previti, O. Harari, I. Zwir, and C. del Val, "Profile analysis and prediction of tissue-specific CpG island methylation classes," *BMC Bioinf.*, vol. 10, p. 116, Apr. 2009.

[23] J. Zhuang, M. Widschwendter, and A. E. Teschendorff, "A comparison of feature selection and classification methods in DNA methylation studies using the Illumina Infinium platform," *BMC Bioinf.*, vol. 13, p. 59, Apr. 2012.

[24] R. Das *et al.*, "Computational prediction of methylation status in human genomic sequences," *Proc. Nat. Acad. Sci. USA*, vol. 103, no. 28, pp. 10713–10716, 2006.

[25] B. Baur and S. Bozdag, "A feature selection algorithm to compute gene centric methylation from probe level methylation data," *PLoS ONE*, vol. 11, no. 2, p. e0148977, 2016.

[26] W. Zhou, C. Zhou, G. Liu, and H. Zhu, "Feature selection for microarray data analysis using mutual information and rough set theory," in *Proc. IFIP Int. Fed. Inf. Process., Artif. Intell. Appl. Innov.*, vol. 204, I. Maglogiannis, K. Karpouzis, and M. Bramer, Eds. Boston, MA, USA: Springer, 2006, pp. 492–499.

[27] M. Nayyeri and H. S. Noghabi, "Cancer classification by correntropy-based sparse compact incremental learning machine," *Gene Rep.*, vol. 3, pp. 31–38, Jun. 2016.

[28] B. T. Moghadam, M. Dabrowski, B. Kaminska, M. G. Grabherr, and J. Komorowski, "Combinatorial identification of DNA methylation patterns over age in the human brain," *BMC Bioinf.*, vol. 17, p. 393, Sep. 2016.

[29] S. Kurdyukov and M. Bullock, "DNA methylation analysis: Choosing the right method," *Biology*, vol. 5, no. 1, p. 3, 2016.

[30] Z. M. Hira, D. F. Gillies, and E Curry, "Improving classification accuracy of response in leukaemia treatment using feature selection over pathway segmentation," Dept. Comput., Imperial College London, London, U.K., Tech. Rep. 2014/8, 2014, http://www.doc.ic.ac.uk/research/technicalreports/2014/DTR14-8.pdf

[31] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," in *Proc. IEEE Conf. Comput. Syst. Bioinf.*, Aug. 2003, pp. 523–528.

[32] M. List, A. C. Hauschild, Q. Tan, T. A. Kruse, J. Baumbach, and R. Batra, "Classification of breast cancer subtypes by combining gene expression and DNA methylation data," *J. Integr. Bioinf.*, vol. 11, no. 2, pp. 1–14, 2014.

[33] V. Jain, W. Zhou, and Y. Men, "Machine learning classification of kidney and lung cancer types," Stanford Univ., Stanford, CA, USA, Tech. Rep., 2013. [Online]. Available: http://cs229.stanford.edu/proj2013/JainMenZhou-MachineLearningClassificationofKidneyandLungCancerTypes.pdf

[34] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[35] J. Li, H. Su, H. Chen, and B. W. Futscher, "Optimal search-based gene subset selection for gene array cancer classification," *IEEE Trans. Inf. Technol. Biomed.*, vol. 11, no. 4, pp. 398–405, Jul. 2007.

[36] X. Li and M. Yin, "Multiobjective binary biogeography based optimization for feature selection using gene expression data," *IEEE Trans. Nanobiosci.*, vol. 12, no. 4, pp. 343–352, Dec. 2013.

[37] A. A. Raweh, M. Nassef, and A. Badr, "A hybridized feature selection and extraction approach for enhancing cancer prediction based on DNA methylation," *IEEE Access*, vol. 6, pp. 15212–15223, 2018.

[38] *RnBeads*. Accessed: Jul. 10, 2018. [Online]. Available http://rnbeads.mpi-inf.mpg.de/

[39] C. C. Aggarwal, *Data Mining: The Textbook*. Cham, Switzerland: Springer, 2015.

[40] X. Wu *et al.*, "Top 10 algorithms in data mining," *Knowl. Inf. Syst.*, vol. 14, no. 1, pp. 1–37, 2008.

[41] T. K. Ho, "Random decision forests," in *Proc. 3rd Int. Conf. Document Anal. Recognit.*, Montreal, QC, Canada, Aug. 1995, pp. 278–282.

[42] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[43] M. Y. H. Al-Shamri, "User profiling approaches for demographic recommender systems," *Knowl.-Based Syst.*, vol. 100, pp. 175–187, May 2016.

[44] N. Japkowicz and M. Shah, *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge, U.K.: Cambridge Univ. Press, 2011, pp. 74–159, doi: 10.1017/CBO9780511921803.

[45] A. Zheng, *Evaluating Machine Learning Models: A Beginner Guide to Key Concepts and Pitfalls*. Newton, MA, USA: O'Reilly Media, 2015, pp. 7–12.

[46] V. B. Canedo, "Novel feature selection methods for high dimensional data," Ph.D. dissertation, Dept. Comput. Sci., Univ. Coruña, Coruña, Spain, Mar. 2014.

[47] I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh, *Feature Extraction: Foundations and Applications*. Cham, Switzerland: Springer, 2006.

[48] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *J. Mach. Learn. Res.*, vol. 5, no. 10, pp. 1205–1224, 2004.

[49] V. Canedo, N. Maroño and A. Betanzos, *Feature Selection for High Dimensional Data*. Berlin, Germany: Springer-Verlag, 2015.

[50] M. W. Berry, S. T. Dumais, and G. W. O'Brien, "Using linear algebra for intelligent information retrieval," *SIAM Rev.*, vol. 37, no. 4, pp. 573–595, 1995.

[51] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Incremental singular value decomposition algorithms for highly scalable recommender systems," in *Proc. 5th Int. Conf. Comput. Inf. Technol. (ICCIT)*, 2002, pp. 27–28.

[52] J. Brownlee. (Oct. 6, 2014). *An Introduction to Feature Selection in Machine Learning Process*. [Online]. Available: http://machinelearningmastery.com/an-introduction-to-feature-selection//

ABDULMAJID F. AL-JUNIAD was born in Taiz, Yemen, in 1975. He received the B.S. degree in control and computer engineering from the University of Technology, Baghdad, Iraq, in 2000, and the M.Sc. and Ph.D. degrees in electrical engineering "computer engineering" from Assiut University, Egypt, in 2007 and 2011, respectively. From 2011 to 2016, he was an Assistant Professor with Electrical Engineering Department, Faculty of Engineering and Architecture, Ibb University, Ibb, Yemen. He is currently an Assistant Professor with Computer Engineering Department, Faculty of Computer Science, KKU University, Abha, Saudi Arabia. His research interests include high performance computation, parallel processing, matrix/vector computation, FPGA/SystemC implementation, multi-core/many-core processors, image processing, image transmission, recommender systems, and artificial intelligence.

TALAL S. QAID was born in Taiz, Yemen, in 1968. He received the B.Sc. degree from Al-Mustansiriya University, Iraq, in 1993, the M.Sc. degree from Pune University, India, in 2000, and the Ph.D. degree from Cairo University, Egypt, in 2009, all in computer science. From 2009 to 2014, he was an Assistant Professor with Computer Science Department, Faculty of Computer Science and Engineering, Hodeidah University, Hodeidah, Yemen. Since 2014, he has been an Assistant Professor with Computer Science Department, Faculty of Computer Science, King Khalid University, Abha, Saudi Arabia. His research interests include recommender systems, data mining, and artificial intelligence.

MOHAMMAD YAHYA H. AL-SHAMRI (M'16) received the B.Sc. degree (Hons.) in electrical engineering from Al-Mustansiriya University, Baghdad, Iraq, in 1999, and the M.Tech. and Ph.D. degrees from the School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi, India, in 2005 and 2008, respectively. He is an Associate Professor with Electrical Engineering Department, Faculty of Engineering and Architecture, Ibb University, Yemen. He is currently with Computer Engineering Department, College of Computer Science, King Khalid University, Abha, Saudi Arabia. He has authored or co-authored many research papers in reputable international journals. His research interests are Web personalization, recommender systems, and bioinformatics.

MAHDI H. A. AHMED received the B.S. and M.S. degrees in electronic engineering and the Ph.D. degree in tele-informatics from Gdansk Technical University, Gdansk, Poland, in 1995 and 2001, respectively. From 2001 to 2004, he was an Assistant Professor at Hodeidah University, Hodeidah, Yemen, from 2004 to 2010, he was with Taiz University, Yemen, and since 2010 he has been on-leave, as an Associate Professor at King Khalid University, Abha, Saudi Arabia. His research interests include wireless networks, routing and data link layer protocols, and recommender systems.

ABEER A. RAWEH was born in Hodeidah, Yemen, in 1981. She received the B.Sc. degree in computer science from Hodeidah University, Hodeidah, in 2004, and the M.Sc. degree in computer science from Cairo University, Egypt, in 2012, where she is currently pursuing the Ph.D. degree in computer science. From 2012 to 2014, she was a Lecturer with Computer Science Department, Faculty of Computer Science and Engineering, Hodeidah University. Since 2014, she has been a Lecturer with Computer Science Department, Faculty of Computer Science, King Khalid University, Abha, Saudi Arabia. Her research interests include recommender systems, data mining, and artificial intelligence.