IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 23, 2022, accepted April 27, 2022, date of publication May 11, 2022, date of current version May 23, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3174260

DNA Methylation Prediction Using Reduced Features Obtained via Gappy Pair Kernel and Partial Least Square

SAJID SHAH[1,2], ALTAF UR RAHMAN[2], SAIMA JABEEN[3], AHMAD KHAN[2], FIAZ GUL KHAN[2], AND MOHAMMED ELAFFENDI[1]

[1]EIAS Laboratory, College of Computer and Information Sciences, Prince Sultan University, Riyadh 11586, Saudi Arabia
[2]Department of Computer Science, COMSATS University Islamabad, Khyber Pakhtunkhwa 22060, Pakistan
[3]Department of IT and Computer Science, Pak-Austria Fachhochschule: Institute of Applied Sciences and Technology, Mang, Haripur, KPK 22620, Pakistan

Corresponding author: Saima Jabeen (saima.jabeen@fecid.paf-iast.edu.pk)

This work was supported by the EIAS Laboratory, CCIS, Prince Sultan University, Riyadh, Saudi Arabia.

ABSTRACT It is critical to correctly identify DNA methylation because it has been linked to a variety of human disorders, particularly cancer. DNA methylation is an epigenetic process that allows cells to alter gene expression. This work deals with a type of DNA methylation called 5-methyl cytosine (m5c), in which the methyl group (CH_3) is attached to the 5^{th} carbon of cytosine. The performances of different machine learning algorithms used for methylation identification are greatly degraded due to poor representation of input sequential data. In the current work, we have proposed a classification model that is based on the extraction of high differentiating features from the sample sequences using gappy pair kernel. Increasing the number of features to better represent a sequence leads to the curse of dimensionality, which is handled by a dimensionality reduction technique called PLS (Partial Least Square). The obtained features are then subjected to multiple classifiers to test the discriminating power of these features. Results are computed for cross species i.e human and mouse, to check the robustness of our proposed model. Finally, the obtained results are compared in terms of sensitivity, specificity, and accuracy with the state-of-the-art approaches. Our proposed approach has outperformed state-of-the-art techniques in all three metrics for both datasets. For research community to test our technique, we have uploaded our code on github (https://github.com/sajidshahbs/gappypairKernel_Rcode).

INDEX TERMS Cross species, DNA methylation, epigenetic modification, feature reduction, gappy pair kernel, linear discrimitive analaysis(LDA), m5c and m6A, partial least square (PLS), SVM.

I. INTRODUCTION

The field of of epigenetics has gained popularity among researchers in the last decade. The term *epigenetics* is used to study a variety of heritable and stable chromatin modifications in the gene expression rather than the primary DNA sequence [2]–[4]. Various kinds of epigenetic features are called *marks*. These marks include histone proteins post-translational modification, DNA methylation, chromatin organization, and non-coding regulatory RNA, etc [5]. When methyl group (CH_3) is attached to the 5^{th} carbon of cytosine, it forms 5-methyl cytosine (m5c) while in case of m6A the methyl group is attached to the nitrogen 6^{th} position of adenosine, in DNA methylation [6]. Some DNA regions, called CpG island, have a high percentage of phosphate bonded Cytosine and Guanine. Twenty-eight million CpG sites have been discovered in the human genome in which 60% to 80% are methylated [7]. DNA methylation contributes to a diverse range of biological procedures, e.g., stable transcriptional gene silencing, X inactivation [8] and genomic imprinting [9]. It is witnessed that DNA methylation also performs a vital role in sustaining cellular function, development of autoimmunity, keeping the genomic stability, and ageing [10], [11]. Such chemical modifications of the bases affect cell events responsible for gene silencing which take part in many diseases such as cancer [12]. Therefore, the prediction of DNA methylation has great importance.

The associate editor coordinating the review of this manuscript and approving it for publication was Yu-Da Lin.

Various experimental approaches have been adopted to analyze DNA methylation. These methods include LC-MS/MS (Liquid Chromatography coupled with tandem mass spectrometry), RFLP (Restriction Fragment Length Polymorphism), HPLC-UV (High-Performance Liquid Chromatography-Ultraviolet), LUMA (LUminometric Methylation Assay) and digestion based as says [13]. The experimental methods are costly and tedious. Furthermore, the advent of the next-generation sequencer has shifted the bottleneck of genomic studies towards the computational epigenetics paradigm. Thus, the production of a huge amount of experimental data has allowed the biologist to design computational methods for effective analysis.

For DNA methylation identification, the computational methods are considered an alternative to lab experimental procedures. Generally, machine learning approaches consist of three basic steps: data collection, features extraction and applying classification model to address any machine learning problem [14]. Experimentally obtained data such as microarrays impose restrictions, due to which researchers initially focused on specific regions (e.g. CGIs) to get satisfactory results [15]–[19]. The analysis of the common regions such as RRBS (Reduced Representation Bisulfite Sequencing) and WGBS (Whole Genome Bisulfite Sequencing), reduces the model performance [20]. The reason is obvious because GC content in positive samples is high while lower in negative samples. Therefore, it is not difficult to distinguish methylated and unmethylated CpGs by utilizing only GC content as a feature. The role and significance of GC content reduces in common regions. To cope with this issue, more complex and discriminative features are required to improve the goodness of the predictions. Feature extraction is considered one of the most important and basic steps for accurate class prediction in machine learning. In other words, high-quality features are key to improve the accuracy of machine learning models. Therefore, the extracted features should be distinct, descriptive and discriminative to help the model to generalize the learning process. Normally, DNA sequences are encoded as numerical values. The feature extraction models are either completely or partially ignore the sequence order information [21]–[24] which leads to low accuracy. Different approaches have been used for feature extraction to study methylation. For example, DNA composition [15], [17], [25], [26], pseudo trinucleotide composition (PseTNC) [21], [27]–[30], predicted DNA structure [15], [31], single nucleotide polymorphisms (SNPs) [15], TFBSs (Transcription Factor Binding Site) [15], [31], histone modifications [15], [31], neighboring CpG site methylation status and distance [31], are some of them.

It is also worth mentioning here that ''Pse-in-One'' is a web-server, which has been used to extract different kinds of features from protein or DNA sequences [32].

Expressing a biological sequence with a vector or discrete model while preserving key pattern characteristic or sequence order information, is the most difficult yet important problems in the field of computational biology. All available machine-learning algorithms such as SVM [33], [34], Nearest Neighbor (NN) [33], Covariance Discriminant (CD) algorithm [35], [36], and Optimization algorithm [37] are capable to work with vectors as discussed in a survey [38]. The problem in defining vector as a discrete model is that sequence-pattern information may lose completely. To overcome this problem of losing sequence-pattern information completely for proteins PseAAC [39] or the pseudo amino acid composition [40] was proposed. Almost all areas of computational proteomics (see, e.g., [41], [42] as well as a number of other papers cited in [43], [44] have made use of the proposed Chou's PseAAC. Due to its wide usage, some popular freely available soft-wares, named as 'PseAAC' [45], ''PseAAC-Builder'' [46], ''propy'' [47], and ''PseAAC-General'' [48] have been developed. The first one is for Chou's general PseAAC [14] which not only includes higher level feature vectors such as ''Sequential Evolution'' or ''PSSM'' mode (see Eqs.13-14 of [14]), ''Functional Domain'' mode (see Eqs.9-10 of [14]), and ''Gene Ontology'' mode (see Eqs.11-12 of [14]) but also all the special modes of feature vectors for proteins while the later three generate different modes of Chou's special PseAAC [49]. The effectiveness of PseAAC for peptide/protein sequences prompted the concept of PseKNC (Pseudo K-tuple Nucleotide Composition) [27] to generate different feature vectors for RNA/DNA sequences [30], [50], [51]. It is worth mentioning that a powerful web-server ''Pse-in-One'' [32] and its next version 'Pse-in-One2.0' were developed for research community to generate their required feature matrices of RNA/DNA sequences [52].

The final major step of a machine learning prediction system is the selection of a classification model. The commonly chosen classification models are: support vector machine (SVM) [7], [15]–[17], [20], [26], [53], random forest (RF) [18], [31], stacked denoising autoencoders (SDA) and naive Bayes (NB) [54]. The majority of the researchers have used the SVM classifier due to its strength in classification and ability to deal with different data types. Some researchers have shown the superiority of the SDA method from deep learning and RF over SVM for predicting DNA methylation [31]. Similar work is carried out using the Deep CpG approach, based on neural networks for predicting methylation states in a single cell [55]. A systematic review in [56] discusses in detail DNA methylation databases, its relationship with different diseases and the machine learning algorithms used for its identification. Some researchers use DNA methylation and machine learning learning to predict human age [57], Parkinsonś Disease [58].

In our proposed technique, we have accommodated sequence order information very effectively, with the help of gappy pair kernel [59], [60]. In other words, our goal is to increase the accuracy of our predicting model with the effective use of sequence order information. The proposed technique takes into account the features count as well as spatial configuration information, like features position in the sequence. Gappy pair kernel is used for the first time, up to

the best of our knowledge, to identify methylation. As mentioned in [59], position information has great importance in finding the similarity of sequences when the underlying transformations are complex. When the number of representations (features) is increased, the issue of high dimensionality arises, which has been tackled by using PLS (Partial Least Square). Finally, the reduced features are passed to multiple classifiers to investigate the discriminative power of these features. We have computed the results for cross-species to test the stability and robustness of the extracted features. The results show that our proposed method has outperformed state-of-the-art techniques for both datasets, i.e. human and mouse, in terms of accuracy, sensitivity and specificity.

The rest of this paper is organized in light of Chou's five-steps rule. The papers to develop a new sequence analyzing method or statistical predictor based on Chou's five-steps rule are expected to have clear logic, transparency in operation(s), ease to regenerate the reported results, availability of high potential guidelines to stimulate other sequence-analyzing methods, and user-friendly interface. The significance of Chou's five-steps rule is discussed in a series of latest publication [61]–[66] and comprehensively summarized by review articles [14], [67], [68]. To develop an effective predictor for a biological system, Chou's five-steps rule is useful to be followed. These steps include: (1) To construct a valid benchmark dataset for training and testing the model. Section 2 discusses the construction of benchmark dataset used in the proposed work. (2) sample sequence representation with a clear mathematical formulation capable to demonstrate the inherent features of the sample sequences. The extracted features are desired to have high correlation between samples and target variable. Feature formulation and extraction is discussed in section II. (3) Employing a powerful machine learning algorithm with high discriminating (classification) power. Classifiers are discussed in section II. (4) Effectively design crossvalidation tests to objectively asses the goodness of the underlying model. Section III discusses evaluation of the model and obtained results. (5) providing a web-server that is easily accessible to the research community. The need of a web-server is mentioned in a series of recent publications [69], [70]), can significantly enhance their impact [38], [67], and driving medicinal chemistry into an unprecedented revolution [43]. We shall put efforts in our future work to provide a web-server for displaying the findings that can be manipulated by users according to their need. For the moment, we have provided our code in github for the ease of research community to regenerate the results. The Github link is: (https://github.com/sajidshahbs/gappypairKernel Rcode). Section IV concludes the proposed work.

II. MATERIAL AND METHODS

A. BENCHMARK DATASET

The main focus of the current work is Human DNA methylation. The DNA of mouse is also used to check the robustness of the proposed work.

A valid benchmark dataset is important to learn a statistical predictor for DNA methylation site prediction. The human DNA dataset [21] along with mouse DNA dataset are used. The $CD-HIT$ tool is used to remove the redundant samples having 70% or above similarity [54]. The human dataset consists of 2426 samples out of which 787 are methylated and 1639 are non-methylated while the mouse dataset consists of 3864 examples out of which 1934 are methylated and 1934 are non-methylated. We have adopted the same procedure as followed by [71] for mouse dataset preparation. For the dataset construction, $41nt$ is used as a sequence length centered at Cytosine. Let S^+ and S^- are the sets of positively and negatively methylated samples respectively, which collectively form the whole dataset S as:

$$S = S^+ \cup S^- \tag{1}$$

B. FEATURE EXTRACTION

Let (X, Y) represents a dataset, where $X = [\boldsymbol{x^1}, \boldsymbol{x^2}, \cdots, \boldsymbol{x^n}]$ contains n number of DNA sequences while $Y = [y^1, y^2, \cdots y^n]$ contains their corresponding labels such that $y^i \in \{0, 1\}$. Thus, a particular DNA sequence $\boldsymbol{x^i}$ containing l nucleotides is given by:

$$\boldsymbol{x^i} = [x_1^i x_2^i x_3^i \ldots x_l^i] \tag{2}$$

where x_j^i represents the j^{th} nucleotide such that $x_j^i \in \{A, C, G, T\}$. For 100 nucleotides of DNA sequence, the total possible combinations are $4^{100} = 1.6065 \times 10^{60}$ [27]. In reality, DNA sequence contains more than 100 nucleotides. Thus, different combinations become enormously large.

In DNA sequence classification, the order of different nucleotides plays an important role. To effectively encode the sequence order information, we have used the gappy pair kernels.

Let $\phi^k(\boldsymbol{x})$ be a feature vector which encodes the frequencies of different sub-strings (k-mers) that are separated by at least m irrelevant positions (m nucleotides). Let u and v be the two k-mers of length k having up to m irrelevant positions between them:

$$\phi^k(\boldsymbol{x}) = \sum_{\forall u,v \in \{A,C,G,T\}^*} I(1)[u \underbrace{\ldots\ldots\ldots}_{0 \le i \le m} v] \tag{3}$$

where $I(1)$ is the indicator function and $\{A, C, G, T\}^*$ represents a set which contains all possible strings of the given nucleotides. The size of the feature vector $\phi^k(\cdot)$ can be calculated as [72]:

$$|\phi^k(\boldsymbol{x})| = (m+1)|A|^{2k} \tag{4}$$

where $|A|$ is the length of the alphabets set (4 for DNA and RNA while 21 for Protein). For example, if $k = 1$ and $m = 21$ then, the length of the resulting feature vector will be 352. Increasing the values of either k or m or both will increase the dimensionality.

The word k-mer or motif is very popular term in computational genomics or sequence analysis. To encode a biological sequence into numerical form, k-mer composition is a very

famous encoding technique. The major issue with k-mer composition is that the order information is completely or partially lost. Order information is very important from the biological aspect. Even if we change the order of only two nucleotides in a sequence, its biological meaning may change completely. By increasing the length of the k-mer we get more order information but with the cost of high dimensionality and vice versa.

Gappy pair kernel basically incorporate more order information while encoding the sequences into numerical form. The parameter k maintain the order with in the motif (k-mer) while m maintain the order between the motifs (k-mers). In other words, k is responsible for maintaining local order (order with in motif) while m is responsible for maintaining global order (order between motifs). Therefore, the values of both k and m have biological significance as well.

C. DIMENSIONALITY REDUCTION

The curse of dimensionality is a well-known phenomenon in machine learning. The high dimensionality of the input dataset often reduces the performance of the underlying classification model due to multiple reasons such as correlated features, inclusion of noise, etc. To cope with this issue, two categories of dimensionality reduction techniques can be used such as supervised (e.g. PLS: Partial Least Squares) and unsupervised techniques (e.g., PCA: Principal Component Analysis). PLS has edge over PCA in terms of the generated results [73]–[75]. A concise discussion about dimensionality reduction is presented in [76].

To reduce the dimensionality, a normalized similarity matrix $\Phi \in \Re^{n\times n}$ is constructed:

$$\Phi = \left[\frac{\phi^k(\boldsymbol{x^i})\phi^k(\boldsymbol{x^j})^T}{||\phi^k(\boldsymbol{x^i})|| \cdot ||\phi^k(\boldsymbol{x^j})||}\right], \quad 1 \le i, j \le n \tag{5}$$

where n shows the number of sample sequences. For further reducing the dimensionality, the PLS algorithm [77] is employed. The generated matrix Φ is still memory intensive. Its space complexity is $O(n^2)$.

Being a supervised method, PLS uses the input features and independent variables to construct latent predictors. It performs better compared to other reduction techniques when the number of predictors (input features) is very large than the number of observations.

By considering the cosine similarity matrix $\Phi \in \Re^{n\times n}$ and the corresponding class labels Y, a factor score matrix $T \in \Re^{n\times r}$, such that $n > r$, is constructed as:

$$T = \Phi W \tag{6}$$

where $W \in \Re^{n\times r}$ is a weight matrix which reflects the covariance structure between the input features Φ and output variables Y. For the estimation of weight matrix W, the SIMPLS algorithm [77] is used.

D. CLASSIFIERS

We have selected five most popular classification models i.e. LDA (Linear Discriminative Analysis), RF (Random Forest), NNET (Neural Network), KNN (K-Nearest Neighbors) and SVM (Support Vector Machine) in our work. The strengths and weaknesses of these classifiers are well explained in the literature, that is why we are not going to explain them here. Analyzing complex relations therein becomes far more easy if they are represented graphically to study medical and biological systems as described in eight foundational papers from the then Chairperson of Nobel Prize Committee Sture Forsen (see, e.g., [78], [79]), and a number of papers cited in a comprehensive review [80], as well as various other papers (see, e.g., [81]. Figure 1 is shown to describe the proposed methodology.

Algorithm 1 SIMPLS Algorithm [77]

Require: Mean centered Φ and Y

1: Initialize: $A_0 = \Phi^T Y$, $M_0 = \Phi^T \Phi$, $C_0 = I$
2: **for** h=0 *to* r **do**
3: compute q_h, the dominant eigenvector of $A_h^T A_h$
4: $w_h = A_h q_h$, $c_h = w_h^T M_h w_h$, $w_h = \frac{w_h}{\sqrt{c_h}}$ and store w_h into W as a column
5: $q_h = A_h^T w_h$, and store q_h into Q as a column
6: $v_h = C_h p_h$, $v_h = \frac{v_h}{||v_h||}$
7: $C_{h+1} = C_h - v_h v_h^T$, $M_{h+1} = M_h - p_h p_h^T$
8: $A_{h+1} = C_h A_h$
9: **end for**

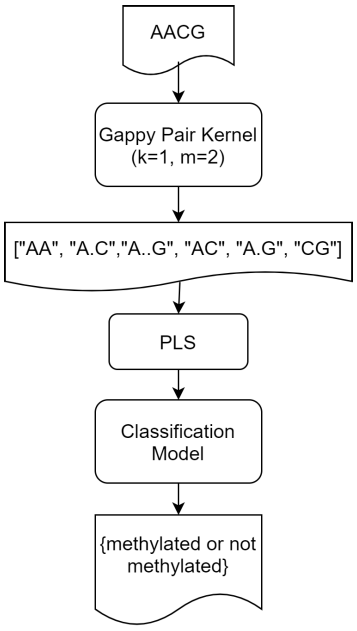


FIGURE 1. Proposed model.

III. RESULTS AND DISCUSSION

Our current work deals with the binary classification where a sequence is either methylated (positive class) or unmethylated (negative class). The performance of our proposed predictor is evaluated using sensitivity (S_n), specificity (S_p) and accuracy (Acc) metrics. Sensitivity or recall (true positive rate) is the ratio of actual positive examples that are

correctly identified while specificity (true negative rate) is the ratio of actual negative examples that are correctly identified. Accuracy is the overall accurate prediction of the samples (both positive and negative) in a dataset. Using these three metrics, the obtained results are compared with the state-of-the-art predictors. Our main goal is to achieve the highest accuracy, with stable values of sensitivity and specificity. The experiments were carried out for both datasets i.e. human and mouse DNA. Accuracy, sensitivity and specificity are formulated as:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$S_n = \frac{TP}{TP + FN} \tag{8}$$

$$S_p = \frac{TN}{TN + FP} \tag{9}$$

where TP represents true positive, TN is for true negative, FP is for false positive (non-methylated sample predicted as methylated sample) and FN is for false negative (methylated sample predicted as non-methylated sample). The ranges of the above formulas are between 0 and 1.

In literature, three popular techniques are used to objectively evaluate the goodness of a predictor: (1) The independent dataset test, (2) The k-fold cross validation (3) The jackknife test.

As mentioned in [14], the independent dataset technique faces the issue of ''memory effect'' or bias in test set selection, therefore, we have avoided to use it.

K-Fold cross-validation splits the dataset into k partitions of nearly equal samples in each partition. One partition referred to a fold, is used for testing and the rest of (k-1) folds are considered for training. This procedure is repeated k times. When all the iterations of the k-fold cross-validation are performed, then the averaged performance measure is calculated to get a generalized performance estimation. In k-fold cross validation, we have two extremes and a middle path. On one extreme, we have Leave One Out Cross Validation (LOOCV), in which one example is used for testing per iteration. This procedure is repeated n times where n is the number of examples. We can call it n-fold cross validation as well. In this case, we have high variance, and the system (model) is normally over fitted. On the other extreme, we have a two fold cross validation in which 50% of the data is used in training while the other 50% is used as testing data. In this case, we have high bias, and the model is normally under fitted. To mitigate the issue of underfitting and overfitting, we have adopted the middle way. In literature, it is a famous topic called the bias-variance trade off. A substantial amount of literature is dedicated to discuss bias-variance trade off. Interested readers can study [82]–[85] etc.

Cross-validation is an effective technique, especially in cases where we need to overcome the problem of overfitting [86]. The most popular k-fold cross-validation is 5-fold so we have adopted it in our work. We can represent 5-fold in terms of ratio as: 80%: 20%, where 80% of the data is dedicated to training and 20% to testing. Another reason for selecting 5-fold cross validation is to make our results compatible with the state of the art. Although, we have created a trade off between bias and variance yet we do not reject the possibility of over fitting because in ideal cases, the testing dataset should not be used in training [87]. Also, we have performed Leave One Out Cross Validation (LOOCV). The jackknife method is similar to LOOCV that is why we have not used it.

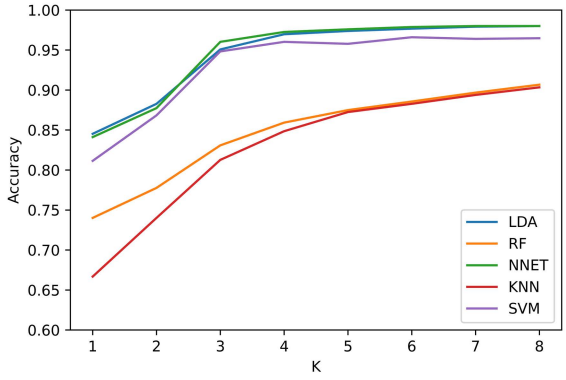
A. EXPERIMENTAL DESIGN

After feature extraction and dimensionality reduction, we have used five classifiers as mentioned in section II-D. It should be noted that the tuning of parameters made for participating classification models is not presented here.

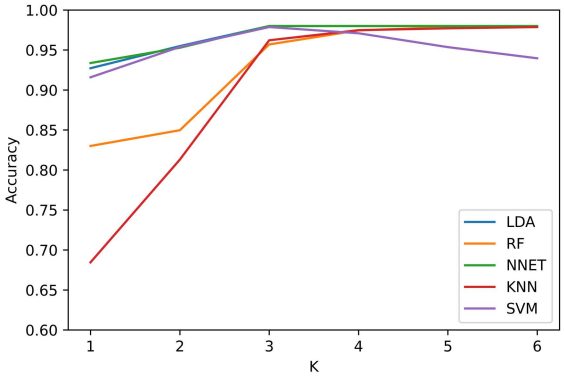
As discussed earlier, we have used a gappy pair kernel for feature extraction. First, we have evaluated the significance of feature extraction with gappy kernel using different combinations of its two parameters; i.e., m and k. Increasing the values of either m or k increases the dimensionality of the input data. On the other hand, increasing the values of m and k preserve sequence order information. In other words, bigger the values of either m, k or both, the sequence order information is preserved. As it is stated earlier that the feature extraction methods either completely or partially ignore this sequence order information. We expect that by preserving order information the performance of the classifiers will enhance. To find suitable values for both m and k and to establish the importance of sequence order information, we performed multiple experiments. First we set $m = 23$ and the latent components for PLS (Partial Least Square) were set to 110 while the value of k was ranging from 1 to 8 for the human dataset while it was from 1 to 6 for mouse dataset. The obtained results are shown in figure 2 where we can see that increasing the value of k improve accuracy of all classifiers. We have stopped increasing the value of k when the accuracies of LDA and NNET reached to maximum i.e., 98%. The accuracies of LDA and NNET reaches to 98% when k was 8 for the human dataset while k was 3 for mouse dataset. Surprisingly the accuracy of SVM decreases when k was beyond 3 for mouse dataset. From these results we decided to set the values of k to 8 and 4 for the human dataset and mouse dataset correspondingly to generate further results.

Gappy pair kernel has another parameter m which is responsible for preserving sequence order information because it deals with irrelevant positions (which creates the gape or tells how many terms will be skipped). We have generated results by setting k to and 4 and the number of latent components to 110 while the value of m was ranging from 10 to 27. The obtained results are shown in figure 3.

Now, it is cleared from both figures 2 and 3 that increasing the values of m and k we are getting better results which means that the order information is playing its role in accordance with our expectation.

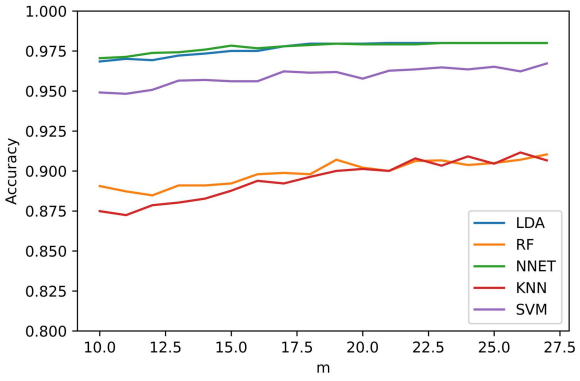


(a) Human

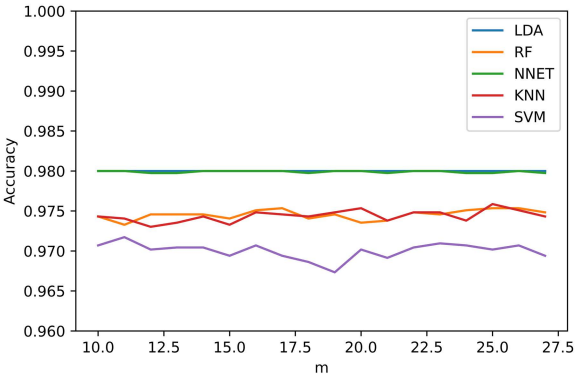


(b) Mouse

FIGURE 2. The impact of K over performances.



(a) Human



(b) Mouse

FIGURE 3. The impact of M over performances.

To tune the number of latent components of PLS to get the optimal dimensionality without compromising the performance, we have performed a set of experiments by setting $m = 23$ for both datasets while keeping $k = 8$ for human and $k = 4$ for mouse dataset. The obtained results are shown in figure 5. The performance of LDA, NNET, and SVM increase while increasing the number of latent components for the human dataset. LDA and NNET reach to 98% for latent components 101 and above. In the case of the mouse dataset, the increase in performance reaches to maximum very early. The best results of all five classifiers with different values of parameters are shown in table 1 and 2 for both datasets. The ROC of both dataset for LDA classifier is given in figure 4.

We performed experiments using 5-fold cross-validation as well as LOOCV (Leave One Out Cross Validation) to investigate the biasness in the results. We achieved almost the same results for both types of cross-validations.

B. COMPARISON WITH EXISTING PREDICTOR(S)

The obtained results are compared with the state-of-the-art techniques. For unbiased and compatible comparison, we selected the techniques from the literature which are using the same datasets that we have used. Furthermore, the same cross-validation technique is adopted for fair comparative analysis. Note that, the results of competitors are directly taken from the corresponding articles. We compared our results in terms of accuracy, sensitivity, and specificity.

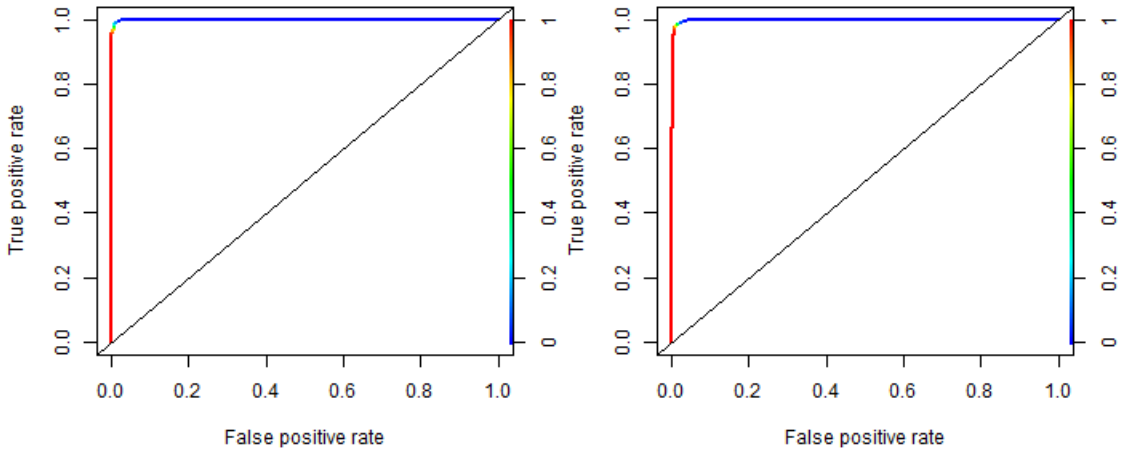
First we discuss the results for the human dataset. The LDA and NNET achieve the 98% accuracies and become the best predictors in our case while the KNN achieves the lowest accuracy of 89.63% and becomes the weakest predictor. Our weakest predictor (KNN) has outperformed both state-of-the-art techniques in terms of accuracy and sensitivity. Our competitors have used SVM as a classification model. Our proposed technique with SVM as a classifier has outperformed both competitors with a big margin in terms of all the three performance metrics. It means when we are using SVM as a classifier then our proposed technique becomes different from the competitors in terms of feature extraction only. This big boost in the performance shows the effectiveness and discriminative power of our feature extraction. In machine learning literature, KNN is a weaker classifier than SVM in complex situations. Here, even KNN outperformed the competitor predictors which shows the strength of gappy pair kernel followed by PLS for feature extraction.

It is worth mentioning that both of our competitors are using 72 features while we have used 109 features for both KNN and SVM. The time used by PLS is additional

TABLE 1. Best results of all classifiers for human dataset.

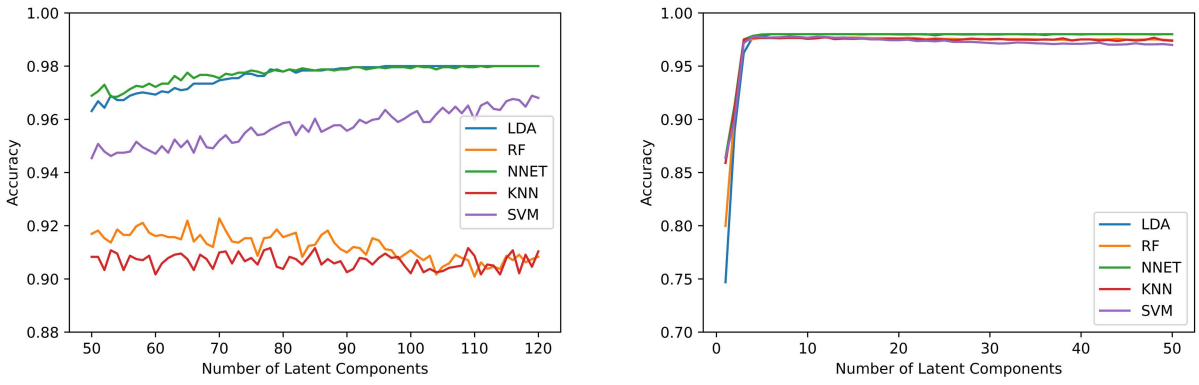
Classifier	$S_n(\%)$	$S_p(\%)$	Acc(%)
iDNA-Methyl [88]	61.25	90.33	77.49
Sequence comp [7]	78.68	78.56	78.62
Results of the Proposed Method			
LDA {nC = 96}	**98.00**	**98.00**	**98.00**
NNET {nC = 101}	**98.00**	**98.00**	**98.00**
SVM {nC =109; C = 01; γ = 0.0065 }	96.96	95.96	96.51
RF {nC = 70; nTree = 02}	95.01	85.54	90.27
KNN { nC = 109; Neighbors=09}	92.70	86.57	89.63
SVM {nC =72; C = 01; γ = 0.011633 }	95.92	95.33	95.62
SVM {nC =50; C = 01; γ = 0.0174 }	95.31	92.66	93.98

- m = 23 and k = 8 for all classifiers
- nC stands for number of components
- C is the regularization parameter and γ is the kernel width for SVM
- nTree is the number of trees in Random Forest



(a) Human (b) Mouse

FIGURE 4. ROC of only LDA classifier for both datasets.



(a) Human (b) Mouse

FIGURE 5. The impact of Latent Components over performances.

overhead. Therefore, we have generated results for SVM using 50 and 72 features. Even, our technique outperformed the competitor predictors using 50 features. The obtained results can be seen in the last two rows of table 1.

To test the stability of the proposed technique we have also used the mouse DNA methylation dataset. The obtained results are shown in table 2. In the case of mouse dataset, again LDA and NNET are the strongest predictors with 98%, accuracy, sensitivity, and specificity while RF (Random Forest) is the weakest with 97.58% accuracy. Since the state-of-the-art has used the Random Forest (RF) and SVM along with other classification models. Thus, we have compared the results of both RF and SVM. Our technique with RF and SVM has outperformed the competing techniques, using RF and SVM.

TABLE 2. Best results of all classifiers for mouse dataset.

Classifier	$S_n(\%)$	$S_p(\%)$	Acc(%)
BayesNet [71]	93.54	98.34	96.04
Random Forest (RF) [71]	93.28	98.34	95.91
SVM [71]	93.28	100	96.73
Results of the Proposed Method			
LDA {nC = 06}	**99.00**	**97.00**	**98.00**
NNET {nC = 06}	**98.00**	**98.00**	**98.00**
SVM {nC = 04; C = 01; γ = 1.896779}	98.00	97.30	97.76
RF {nC = 07; nTree = 02}	97.48	97.69	97.58
KNN {nC = 12; Neighbors=07}	97.53	97.69	97.61

· m = 23 and k = 4 for all classifiers
· nC stands for number of components
· C is the regularization parameter and γ is the kernel width for SVM
· nTree is the number of trees in Random Forest

It is worth discussing that the state-of-the-art methods are using 164 features but we have used merely 4 and 7 features for SVM and RF respectively. Thus, we achieve 41 times lower dimensionality for SVM and 23 times lower for RF.

In a nutshell, the performance of LDA and NNET is remained stable in reduced space for both datasets (cross-species) and achieves remarkable results with 98% accuracy, sensitivity, and specificity.

IV. CONCLUSION

A DNA methylation identification system based on gappy pair kernel and PLS is proposed. Gappy pair kernel and PLS are used for feature extraction and dimension reduction respectively. In this work, we have investigated the significance of maintaining sequence order information in feature extraction and also the importance of dimensionality reduction. Our proposed predictor has outperformed the state-of-the-art techniques in terms of sensitivity, specificity, and accuracy in case of human and mouse DNA methylation. Obtaining 98% performance metrics introduces the possibility of model over fitting. In future, we aim to find whether our model is over fitted or not by rigorous testing using independent dataset test and jackknife testing methods along with cross validation. Furthermore, we aim to test out model on cross species (both animals and plants) to explore the robustness of our model.

ACKNOWLEDGMENT

The authors would like to thank EIAS Data Science Labarotary and Prince Sultan University for their encouragement, support and the facilitation of resources needed and funding to complete this work.

REFERENCES

[1] G. O. Young, "Synthetic structure of industrial plastics," in *Plastics*, vol. 3, 2nd ed., J. Peters, Ed. New York, NY, USA: McGraw-Hill, 1964, pp. 15–64.
[2] E. Jablonka and G. Raz, "Transgenerational epigenetic inheritance: Prevalence, mechanisms, and implications for the study of heredity and evolution," *Quart. Rev. Biol.*, vol. 84, no. 2, pp. 76–131, 2009.
[3] I. M. Fingerman, L. McDaniel, X. Zhang, W. Ratzat, T. Hassan, Z. Jiang, R. F. Cohen, and G. D. Schuler, "NCBI epigenomics: A new public resource for exploring epigenomic data sets," *Nucleic Acids Res.*, vol. 39, pp. D908–D912, Jan. 2011.
[4] S. Berger, T. Kouzarides, R. Shiekhattar, and A. Shilatifard, "An operational definition of epigenetic," *Genes Develop.*, vol. 23, no. 7, pp. 781–783, 2009.
[5] B. E. Bernstein, A. Meissner, and E. S. Lander, "The mammalian epigenome," *Cell*, vol. 128, no. 4, pp. 669–681, Feb. 2007.
[6] L. D. Moore, T. Le, and G. Fan, "DNA methylation and its basic function," *Neuropsychopharmacology*, vol. 38, no. 1, pp. 23–38, Jan. 2013.
[7] C. Wu, S. Yao, X. Li, C. Chen, and X. Hu, "Genome-wide prediction of DNA methylation using DNA composition and sequence complexity in human," *Int. J. Mol. Sci.*, vol. 18, no. 2, p. 420, Feb. 2017.
[8] C. L. Anderson and C. J. Brown, "Variability of X chromosome inactivation: Effect on levels of TIMP1 RNA and role of DNA methylation," *Hum. Genet.*, vol. 110, no. 3, pp. 271–278, Mar. 2002.
[9] M. Li, N. Y. Kim, S. Masuda, and J. C. Belmonte, "Regulation of somatic stem cell function by DNA methylation and genomic imprinting," *Cell Tissue Transplantation Therapy*, vol. 5, p. 19, 2013.
[10] A. R. Elhamamsy, "DNA methylation dynamics in plants and mammals: Overview of regulation and dysregulation," *Cell Biochem. Function*, vol. 34, no. 5, pp. 289–298, Jul. 2016.
[11] J. Su, X. Shao, H. Liu, S. Liu, Q. Wu, and Y. Zhang, "Genome-wide dynamic changes of DNA methylation of repetitive elements in human embryonic stem cells and fetal fibroblasts," *Genomics*, vol. 99, no. 1, pp. 10–17, Jan. 2012.
[12] K. Chiappinelli, P. Strissel, A. Desrichard, H. Li, C. Henke, B. Akman, A. Hein, N. Rote, L. Cope, A. Snyder, V. Makarov, S. Budhu, J. Wolchok, C. Zahnow, T. Mergoub, T. Chan, R. Strick, and S. Baylin, "Abstract B32: Inhibiting DNA methylation causes an interferon response in cancer via dsRNA including endogenous retroviruses," *Cancer Res.*, vol. 76, no. 2, p. B32, 2016.
[13] S. Kurdyukov and M. Bullock, "DNA methylation analysis: Choosing the right method," *Biology*, vol. 5, no. 1, p. 3, Jan. 2016.
[14] K.-C. Chou, "Some remarks on protein attribute prediction and pseudo amino acid composition," *J. Theor. Biol.*, vol. 273, no. 1, pp. 236–247, Mar. 2011.
[15] C. Bock, M. Paulsen, S. Tierling, T. Mikeska, T. Lengauer, and J. Walter, "CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure," *PLoS Genet.*, vol. 2, no. 3, p. e26, Mar. 2006.
[16] S. Fan, M. Q. Zhang, and X. Zhang, "Histone methylation marks play important roles in predicting the methylation status of CpG islands," *Biochem. Biophys. Res. Commun.*, vol. 374, no. 3, pp. 559–564, Sep. 2008.
[17] H. Zheng, H. Wu, J. Li, and S.-W. Jiang, "CpGIMethPred: Computational model for predicting methylation status of CpG islands in human genome," *BMC Med. Genomics*, vol. 6, no. 1, pp. 1–12, Jan. 2013.
[18] 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, no. 1, pp. 1–16, Dec. 2009.
[19] D. Schübeler, "Function and information content of DNA methylation," *Nature*, vol. 517, no. 7534, pp. 321–326, Jan. 2015.
[20] B. Ma, E. H. Wilker, S. A. G. Willis-Owen, H.-M. Byun, K. C. C. Wong, V. Motta, A. A. Baccarelli, J. Schwartz, W. O. C. M. Cookson, K. Khabbaz, M. A. Mittleman, M. F. Moffatt, and L. Liang, "Predicting DNA methylation level across human tissues," *Nucleic Acids Res.*, vol. 42, no. 6, pp. 3515–3528, Apr. 2014.
[21] Z. Liu, X. Xiao, W.-R. Qiu, and K.-C. Chou, "Benchmark data for identifying DNA methylation sites via pseudo trinucleotide composition," *Data Brief*, vol. 4, pp. 87–89, Sep. 2015.
[22] S. J. Cokus, S. Feng, X. Zhang, Z. Chen, B. Merriman, C. D. Haudenschild, S. Pradhan, S. F. Nelson, M. Pellegrini, and S. E. Jacobsen, "Shotgun bisulphite sequencing of the arabidopsis genome reveals DNA methylation patterning," *Nature*, vol. 452, no. 7184, pp. 215–219, Mar. 2008.
[23] Y. Sun, A. K. C. Wong, and M. S. Kamel, "Classification of imbalanced data: A review," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 23, no. 4, pp. 687–719, 2009.
[24] J. Laurikkala, "Improving identification of difficult small classes by balancing class distribution," in *Artificial Intelligence in Medicine*, vol. 2101. Springer 2001, pp. 63–66.
[25] F. Fang, S. Fan, X. Zhang, and M. Q. Zhang, "Predicting methylation status of CpG islands in the human brain," *Bioinformatics*, vol. 22, no. 18, pp. 2204–2209, Sep. 2006.
[26] R. Das, N. Dimitrova, Z. Xuan, R. A. Rollins, F. Haghighi, J. R. Edwards, J. Ju, T. H. Bestor, and M. Q. Zhang, "Computational prediction of methylation status in human genomic sequences," *Proc. Nat. Acad. Sci. USA*, vol. 103, no. 28, pp. 10713–10716, Jul. 2006.

[27] W. Chen, T.-Y. Lei, D.-C. Jin, H. Lin, and K.-C. Chou, "PseKNC: A flexible web server for generating pseudo *K*-tuple nucleotide composition," *Anal. Biochem.*, vol. 456, pp. 53–60, Jul. 2014.

[28] W. Chen, X. Zhang, J. Brooker, H. Lin, L. Zhang, and K.-C. Chou, "PseKNC-general: A cross-platform package for generating various modes of pseudo nucleotide compositions," *Bioinformatics*, vol. 31, no. 1, pp. 119–120, Jan. 2015.

[29] B. Liu, F. Liu, L. Fang, X. Wang, and K.-C. Chou, "RepDNA: A Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects," *Bioinformatics*, vol. 31, no. 8, pp. 1307–1309, Apr. 2015.

[30] W. Chen, H. Lin, and K.-C. Chou, "Pseudo nucleotide composition or PseKNC: An effective formulation for analyzing genomic sequences," *Mol. Biosyst.*, vol. 11, no. 10, pp. 2620–2634, 2015.

[31] W. Zhang, T. D. Spector, P. Deloukas, J. T. Bell, and B. E. Engelhardt, "Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements," *Genome Biol.*, vol. 16, no. 1, p. 19, Dec. 2015.

[32] B. Liu, F. Liu, X. Wang, J. Chen, L. Fang, and K.-C. Chou, "Pse-in-one: A web server for generating various modes of pseudo components of DNA, RNA, and protein sequences," *Nucleic Acids Res.*, vol. 43, no. W1, pp. W65–W71, Jul. 2015.

[33] L. Hu, T. Huang, X. Shi, W.-C. Lu, Y.-D. Cai, and K.-C. Chou, "Predicting functions of proteins in mouse based on weighted protein-protein interaction network and protein hybrid properties," *PLoS ONE*, vol. 6, no. 1, Jan. 2011, Art. no. e14556.

[34] Y.-D. Cai, K.-Y. Feng, W.-C. Lu, and K.-C. Chou, "Using LogitBoost classifier to predict protein structural classes," *J. Theor. Biol.*, vol. 238, no. 1, pp. 172–176, Jan. 2006.

[35] K.-C. Chou and D. W. Elrod, "Bioinformatical analysis of G-protein-coupled receptors," *J. Proteome Res.*, vol. 1, no. 5, pp. 429–433, Oct. 2002.

[36] K.-C. Chou and Y.-D. Cai, "Prediction and classification of protein subcellular location-sequence-order effect and pseudo amino acid composition," *J. Cellular Biochem.*, vol. 90, no. 6, pp. 1250–1260, Dec. 2003.

[37] C.-T. Zhang and K.-C. Chou, "An optimization approach to predicting protein structural class from amino acid composition," *Protein Sci.*, vol. 1, no. 3, pp. 401–408, Mar. 1992.

[38] K.-C. Chou, "Impacts of bioinformatics to medicinal chemistry," *Medicinal Chem.*, vol. 11, no. 3, pp. 218–234, Mar. 2015.

[39] K.-C. Chou, "Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes," *Bioinformatics*, vol. 21, no. 1, pp. 10–19, Jan. 2005.

[40] K. C. Chou, "Prediction of protein cellular attributes using pseudo-amino acid composition," *Proteins Struct. Function Bioinf.*, vol. 43, no. 3, pp. 246–255, 2001.

[41] A. Dehzangi, R. Heffernan, A. Sharma, J. Lyons, K. Paliwal, and A. Sattar, "Gram-positive and gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou's general PseAAC," *J. Theor. Biol.*, vol. 364, pp. 284–294, Jan. 2015.

[42] P. K. Meher, T. K. Sahu, V. Saini, and A. R. Rao, "Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC," *Sci. Rep.*, vol. 7, no. 1, pp. 1–12, Mar. 2017.

[43] K.-C. Chou, "An unprecedented revolution in medicinal chemistry driven by the progress of biological science," *Current Topics Medicinal Chem.*, vol. 17, no. 21, pp. 2337–2358, 2017.

[44] K.-C. Chou, "Progresses in predicting post-translational modification," *Int. J. Peptide Res. Therapeutics*, vol. 26, pp. 873–888, Jul. 2019.

[45] H.-B. Shen and K.-C. Chou, "PseAAC: A flexible web server for generating various kinds of protein pseudo amino acid composition," *Anal. Biochem.*, vol. 373, no. 2, pp. 386–388, Feb. 2008.

[46] P. Du, X. Wang, C. Xu, and Y. Gao, "PseAAC-builder: A cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions," *Anal. Biochem.*, vol. 425, no. 2, pp. 117–119, Jun. 2012.

[47] D.-S. Cao, Q.-S. Xu, and Y.-Z. Liang, "Propy: A tool to generate various modes of Chou's PseAAC," *Bioinformatics*, vol. 29, no. 7, pp. 960–962, Apr. 2013.

[48] P. Du, S. Gu, and Y. Jiao, "PseAAC-general: Fast building various modes of general form of Chou's pseudo-amino acid composition for large-scale protein datasets," *Int. J. Mol. Sci.*, vol. 15, no. 3, pp. 3495–3506, Feb. 2014.

[49] K.-C. Chou, "Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology," *Current Proteomics*, vol. 6, no. 4, pp. 262–274, Dec. 2009.

[50] B. Liu, F. Yang, D.-S. Huang, and K.-C. Chou, "IPromoter-2L: A two-layer predictor for identifying promoters and their types by multi-window-based PseKNC," *Bioinformatics*, vol. 34, no. 1, pp. 33–40, Jan. 2018.

[51] M. Tahir, H. Tayara, and K. T. Chong, "IRNA-PseKNC(2methyl): Identify RNA 2'-O-methylation sites by convolution neural network and Chou's pseudo components," *J. Theor. Biol.*, vol. 465, pp. 1–6, Mar. 2019.

[52] B. Liu, H. Wu, and K.-C. Chou, "Pse-in-one 2.0: An improved package of web servers for generating various modes of pseudo components of DNA, RNA, and protein sequences," *Natural Sci.*, vol. 9, no. 4, p. 67, 2017.

[53] M. Bhasin, H. Zhang, E. L. Reinherz, and P. A. Reche, "Prediction of methylated CpGs in DNA sequences using a support vector machine," *FEBS Lett.*, vol. 579, no. 20, pp. 4302–4308, Aug. 2005.

[54] Y. Wang, T. Liu, D. Xu, H. Shi, C. Zhang, Y.-Y. Mo, and Z. Wang, "Predicting DNA methylation state of CpG dinucleotide using genome topological features and deep networks," *Sci. Rep.*, vol. 6, no. 1, Apr. 2016, Art. no. 19598.

[55] C.-A. Kapourani and G. Sanguinetti, "Higher order methylation features for clustering and prediction in epigenomic studies," *Bioinformatics*, vol. 32, no. 17, pp. i405–i412, Sep. 2016.

[56] C. Ao, L. Gao, and L. Yu, "Research progress in predicting DNA methylation modifications and the relation with human diseases," *Current Medicinal Chem.*, vol. 29, no. 5, pp. 822–836, Feb. 2022.

[57] A. Zaguia, D. Pandey, S. Painuly, S. K. Pal, V. K. Garg, and N. Goel, "DNA methylation biomarkers-based human age prediction using machine learning," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–11, Jan. 2022.

[58] J. Augustine and A. Jereesh, "Blood-based DNA methylation marker identification for Parkinson's disease prediction," in *Proc. Int. Conf. Innov. Comput. Commun.*, 2022, pp. 777–784.

[59] P. Kuksa, P.-H. Huang, and V. Pavlovic, "Fast protein homology and fold detection with sparse spatial sample kernels," in *Proc. 19th Int. Conf. Pattern Recognit.*, Dec. 2008, pp. 1–4.

[60] C. C. Mahrenholz, I. G. Abfalter, U. Bodenhofer, R. Volkmer, and S. Hochreiter, "Complex networks govern coiled-coil oligomerization—Predicting and profiling by means of a machine learning approach," *Mol. Cellular Proteomics*, vol. 10, no. 5, pp. 1–9, 2011.

[61] X. Du, Y. Diao, H. Liu, and S. Li, "MsDBP: Exploring DNA-binding proteins by integrating multiscale sequence information via Chou's five-step rule," *J. Proteome Res.*, vol. 18, no. 8, pp. 3119–3132, 2019.

[62] M. Kabir, S. Ahmad, M. Iqbal, and M. Hayat, "INR-2L: A two-level sequence-based predictor developed via Chou's 5-steps rule and general PseAAC for identifying nuclear receptors and their families," *Genomics*, vol. 112, no. 1, pp. 276–285, Jan. 2020.

[63] K.-C. Chou, "Other mountain stones can attack jade: The 5-steps rule," *Natural Sci.*, vol. 12, no. 3, pp. 59–64, 2020.

[64] K.-C. Chou, "Proposing 5-steps rule is a notable milestone for studying molecular biology," *Natural Sci.*, vol. 12, no. 3, p. 74, 2020.

[65] K.-C. Chou, "Using similarity software to evaluate scientific paper quality is a big mistake," *Natural Sci.*, vol. 12, no. 3, p. 42, 2020.

[66] K.-C. Chou, "The development of Gordon life science institute: Its driving force and accomplishments," *Natural Sci.*, vol. 12, no. 4, pp. 202–217, 2020.

[67] K.-C. Chou, "Advances in predicting subcellular localization of multi-label proteins and its implication for developing multi-target drugs," *Current Medicinal Chem.*, vol. 26, no. 26, pp. 4918–4943, Oct. 2019.

[68] K.-C. Chou, "Impacts of pseudo amino acid components and 5-steps rule to proteomics and proteome analysis," *Current Topics Medicinal Chem.*, vol. 19, no. 25, pp. 2283–2300, Nov. 2019.

[69] X. Cheng, S.-G. Zhao, W.-Z. Lin, X. Xiao, and K.-C. Chou, "PLoc-mAnimal: Predict subcellular localization of animal proteins with both single and multiple sites," *Bioinformatics*, vol. 33, no. 22, pp. 3524–3531, Nov. 2017.

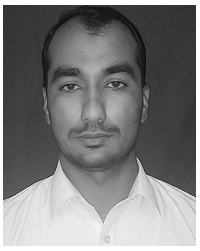
[70] X. Xiao, X. Cheng, G. Chen, Q. Mao, and K.-C. Chou, "PLoc_bal-mGpos: Predict subcellular localization of gram-positive bacterial proteins by quasi-balancing training dataset and PseAAC," *Genomics*, vol. 111, no. 4, pp. 886–892, Jul. 2019.

[71] P. Feng, H. Yang, H. Ding, H. Lin, W. Chen, and K.-C. Chou, "IDNA6 mA-PseKNC: Identifying DNA N6-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC," *Genomics*, vol. 111, no. 1, pp. 96–102, Jan. 2019.

[72] J. Palme, S. Hochreiter, and U. Bodenhofer, "KeBABS: An R package for kernel-based analysis of biological sequences," *Bioinformatics*, vol. 31, no. 15, pp. 2574–2576, Aug. 2015.

[73] S. Wold, A. Ruhe, H. Wold, and W. J. Dunn, "The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses," *SIAM J. Sci. Stat. Comput.*, vol. 5, no. 3, pp. 735–743, Sep. 1984.
[74] H. Wold, "Path models with latent variables: The NIPALS approach," in *Quantitative Sociology*, H. M. Blalock, Ed. Elsevier, 1975, pp. 307–357.
[75] M. Hilario and A. Kalousis, "Approaches to dimensionality reduction in proteomic biomarker studies," *Briefings Bioinf.*, vol. 9, no. 2, pp. 102–118, Sep. 2007.
[76] A. Khan, S. Shah, F. Wahid, F. G. Khan, and S. Jabeen, "Identification of microRNA precursors using reduced and hybrid features," *Mol. BioSyst.*, vol. 13, no. 8, pp. 1640–1645, 2017.
[77] S. de Jong, "SIMPLS: An alternative approach to partial least squares regression," *Chemometric Intell. Lab. Syst.*, vol. 18, no. 3, pp. 251–263, Mar. 1993.
[78] K.-C. Chou and S. Forsén, "Diffusion-controlled effects in reversible enzymatic fast reaction systems–critical spherical shell and proximity rate constant," *Biophys. Chem.*, vol. 12, nos. 3–4, pp. 255–263, Dec. 1980.
[79] T. Li, K. Chou, and S. Forsen, "The flow of substrate molecules in fast enzyme-catalyzed reaction systems," *Chemica Scripta*, vol. 16, no. 5, pp. 192–196, 1980.
[80] G. P. Zhou and M. Deng, "An extension of Chou's graphic rules for deriving enzyme kinetic equations to systems involving parallel reaction pathways," *Biochem. J.*, vol. 222, no. 1, p. 169, 1984.
[81] K.-C. Chou, "Low-frequency collective motion in biomacromolecules and its biological functions," *Biophys. Chem.*, vol. 30, no. 1, pp. 3–48, May 1988.
[82] R. Kohavi and D. H. Wolpert, "Bias plus variance decomposition for zero-one loss functions," in *Proc. ICML*, vol. 96, 1996, pp. 275–283.
[83] B. Neal, "On the bias-variance tradeoff: Textbooks need an update," 2019, *arXiv:1912.08286*.
[84] B. Neal, S. Mittal, A. Baratin, V. Tantia, M. Scicluna, S. Lacoste-Julien, and I. Mitliagkas, "A modern take on the bias-variance tradeoff in neural networks," 2018, *arXiv:1810.08591*.
[85] U. von Luxburg and B. Schkopf, "Statistical learning theory: Models, concepts, and results," in *Handbook of the History of Logic*, vol. 10. Elsevier, 2011, pp. 651–706.
[86] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory undersampling for class-imbalance learning," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 39, no. 2, pp. 539–550, Apr. 2009.
[87] J. Brownlee, "What is the difference between test and validation datasets," *Mach. Learn. Mastery*, vol. 14, 2017.
[88] Z. Liu, X. Xiao, W.-R. Qiu, and K.-C. Chou, "IDNA-methyl: Identifying DNA methylation sites via pseudo trinucleotide composition," *Anal. Biochem.*, vol. 474, pp. 69–77, Apr. 2015.

SAJID SHAH received the M.S. and Ph.D. degrees from the Politecnico di Torino, Italy. He worked as an Assistant Professor with COMSATS University Islamabad, Abbottabad Campus, Pakistan. He is currently a Postdoctoral Researcher with the EIAS Laboratory, Prince Sultan University, Riyadh, Saudi Arabia. His research interests include data mining, text mining, machine learning, bioinformatics, and image processing.

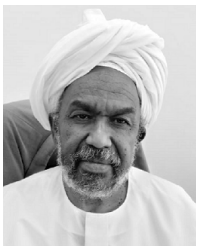


ALTAF UR RAHMAN received the B.S. degree in computer science from S.B.B.U Sheringal, in 2013, and the M.S. degree from COMSATS University Islamabad, in 2017. Since 2017, he has been with KP Elementary and Secondary Education Department as S.S.T-I.T.

SAIMA JABEEN received the Ph.D. degree in computer and control engineering from the Politecnico di Torino, Italy, in 2014. She was an Assistant Professor with the Faculty of Computer Sciences and Engineering, Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, Topi, Swabi, Pakistan, and Prince Sultan University, Riyadh, Saudi Arabia. She also served as an Assistant Professor and the Chairperson of the Department of Computer Science, University of Wah, Wah Cantt, Pakistan. She is currently with the Department of IT and Computer Science, Pak-Austria Fachhochschule: Institute of Applied Sciences and Technology, Mang, Haripur, KPK, Pakistan. Her research interests include data mining, machine learning, NLP, document analysis, and social network analysis.

AHMAD KHAN received the Ph.D. degree from the National University of Computer and Emerging Sciences (FAST-NU), Islamabad, Pakistan, in 2015. He is currently working as an Assistant Professor with the Department of Computer Science, COMSATS University Islamabad (CUI), Abbottabad Campus. His research interests include computer vision, machine learning, and evolutionary algorithms.

FIAZ GUL KHAN received the M.S. and Ph.D. degrees from the Politecnico di Torino Italy, in 2013. He is currently serving as an Assistant Professor with the Department of Computer Science, COMSATS University Islamabad, Abbottabad Campus, Pakistan. His research interests include concurrent computing, machine learning, artificial intelligence, and GPU computing.



MOHAMMED ELAFFENDI is a Professor of computer science with the Department of Computer Science, College of Computer and Information Sciences (CCIS), Prince Sultan University, Riyadh, KSA. His research interests include machine learning, natural language processing, emerging distributed architectures, next-generation computer systems, and complexity science. He is the Founder and the Director of the Center of Excellence in Cybersecurity (CYBEX), the Founder and the Director of EIAS Data Science Research Laboratory, the AIDE to the Rector, the Director of the Institutional Policy, and the Development Unit and a Former Dean of CCIS.