2023 IEEE Symposium Series on Computational Intelligence (SSCI)
Mexico City, Mexico. December 5-8, 2023

A Novel Feature Extraction Approach for the Clustering and Classification of Genome Sequences

Rajesh Dwivedi
Dept. of Computer Science and Engg.
Indian Institute of Technology Indore
rajeshdwivedi@iiti.ac.in

Aruna Tiwari
Dept. of Computer Science and Engg.
Indian Institute of Technology Indore
artiwari@iiti.ac.in

Neha Bharill
Dept. of Computer Science and Engg.
Mahindra University, Hyderabad
neha.bharill@mahindrauniversity.edu.in

Milind Ratnaparkhe
Dept. of Biotechnology
ICAR-IISR Indore
Milind.Ratnaparkhe@icar.gov.in

Abhishek Tripathi
Dept. of Computer Science and Engg.
Indian Institute of Technology Indore
phd2201101002@iiti.ac.in

Preeti Jha
Koneru Lakshmaiah Education Foundation
Bowrampet, Hyderabad, India
Preetijha@klh.edu.in

***Abstract*—Feature extraction is essential in bioinformatics because it transforms genome sequences into the feature vectors required for data mining activities such as classification and clustering. The data mining activities enable us to classify or cluster the newly sequenced genome to the known families. Nowadays, a variety of feature extraction strategies are available for genome data. Nevertheless, several existing algorithms do not extract context-sensitive key properties, also some approaches extract features, which are unable to distinguish between two non-similar sequences. In addition, the efficacy of existing feature extraction techniques is evaluated on either supervised or unsupervised learning models, but not on both. Thus, an efficient feature extraction technique that extracts significantly relevant features from genome sequences is required. In this paper, a novel feature extraction method is proposed that extracts features based on the length of the sequence, the frequency of nucleotide bases, the modified positional sum of nucleotide bases, the distribution of nucleotide bases, and the entropy of the sequence to generate a 14-dimensional fixed-length numeric vector to describe each genome sequence uniquely. By applying extracted features to both supervised and unsupervised machine learning approaches, the performance of the proposed feature extraction method is assessed. The experimental results show that the proposed strategy for clustering and classifying novel genome sequences into recognized genome classes is highly effective and efficient. The same is proven by comparing the proposed method to the standard state-of-the-art method.**

***Index Terms*—Feature extraction, Genome sequences, Clustering, Classification, Single nucleotide polymorphism**

I. Introduction

In bioinformatics, the classification or clustering of biological sequences [1] is a crucial problem. Biologists are frequently interested in determining the family of a newly generated genome sequence. This will allow scientists to investigate the evolution and biological activities of this genome. Biologists typically use sequence alignment to search for sequence similarity and homology in order to classify novel biological sequences into established families/classes. Yet, this method is exceedingly ineffective. One of the biggest problems with using this method, for example, in metagenomics, is that between 25% and 65% of the sequences have no homolog (orphan sequences) in the databases, making these sequences useless for any further analysis [2]. Employing techniques of machine learning is one method for overcoming such problems. Due to the nature of biological sequences, in which features are embedded within the sequence itself, it is not viable to apply the well-known classification methods that are highly effective in real-world data mining applications when applied to relational data. Thus, a method of feature extraction is required to transform biological data into a new format suited for various data mining algorithms [3]. Extracting features from genome sequences [3] allows for the conversion of genome sequences into numerical data, which makes it possible to use data mining techniques to study these sequences.

The set of instructions that comprise an organism is referred to as the Genome, and it is composed of Deoxyribonucleic Acid (DNA) [4]. The four nucleotide bases that makeup DNA are adenine (A), cytosine (C), guanine (G), and thymine (T). A single nucleotide polymorphism (SNP) [3] is a type of mutation/alteration in DNA that can occur when a single nucleotide is added to or removed from the DNA.

This remaining paper continues as follows: Section II reviews the existing literature. Section III introduces the proposed novel 14-dimensional feature extraction technique for the clustering and classification of genome data. Section IV presents the experimental findings on the benchmark and real-life plant genome datasets in comparison with other state-of-the-art approach. Finally, Section V concludes and outlines future research.

978-0-7381-4408-5/23/$31.00 ©2023 IEEE

II. Literature Survey

Marwah A. Helaly et al. [4] proposed a deep learning strategy for the taxonomy classification of bacterial sequence. To describe the genome data, they employed a variety of representations, including one-hot encoding, inter-encoding, and k-mers-based representation. They evaluated their strategy on the 16S rRNA dataset utilizing a deeper convolution neural network (CNN) and obtained an accuracy of 91.7% with a more representative representation and 90.0% with a less figurative representation. The problem with this method is that it can only be used with labeled data. After that, Jasbir Dhaliwal and John Wanger [5] made a new way to extract features for SNPs that are highly expressed. As features, they employed k-mer to describe the SNP sequence. According to them, ideal k-mer and feature size may vary between research problems. They assessed their technique using a multinomial naive bayes on 49 human tissues and obtained optimal k-mer of size 3. One of the biggest problems with using k-mer is that storing an SNP sequence with large-sized k-mers takes a lot of memory. Later on, Preeti Jha et al. [3] presented a feature extraction method named 12d-FV for the SNP sequencing analysis of unlabeled real-world plant genome datasets. To describe an SNP sequence, they employed three sorts of features: frequency, total distance, and nucleotide distribution. They used kernelized scalable random sampling with iterative fuzzy c-means (KSRSIO-FCM) to test their method and evaluated the results using the silhouette index [**?**]. The disadvantage of this strategy is that the total distance and the distribution for each nucleotide may be the same for dissimilar sequences. This technique may be incapable of differentiating between the sequences as a result of this incapability. In addition, the sequence length, which differs between organisms, has not been used in this method. In 2022, Bonidia et al. [6] proposed a new package named MathFeature, for extracting the numerical features from the ribonucleic acid (RNA), DNA and protein sequences. In this package they used 20 numerical feature extraction descriptors based on the numerous numeric mappings, chaos game theory, gemonic signal processing, complex networks and entropy for converting the biological sequences into numerical values. They evaluated their method on eight benchmark datasets and found that MathFeature outperformed competing methods.

According to the aforementioned literature, the majority of feature extraction algorithms are either available for labeled data or unlabeled data, but not for both. In addition, several algorithms do not extract context-based properties. In contrast, some algorithms fail to extract essential features such as length and entropy. To overcome the limitations identified in this study, a novel 14-dimensional feature extraction technique is proposed, which extracts features based on length, frequency, modified position sum, distribution, and entropy to characterize the genome sequence uniquely.

In the further section we will brief about the concept-by-concept analysis of the proposed method, i.e., the five distinct categories of features and their extraction procedure, illustrated with an example and the implementation of proposed approach.

III. Methodology

This section presents a novel 14-dimensional feature extraction technique (14d-FET) that extracts five different types of sequence features, namely sequence length, frequency of nucleotide bases, modified positional sum of nucleotide bases, distribution of nucleotide bases, and sequence entropy. The proposed method also eliminates the disadvantage of having the same positional sum and same distribution in two dissimilar sequences by employing a novel power mechanism. The architecture and pseudo code of of proposed approach is shown in Fig. 1 and Algorithm 1, respectively. The five different types of features are discussed subsequently.

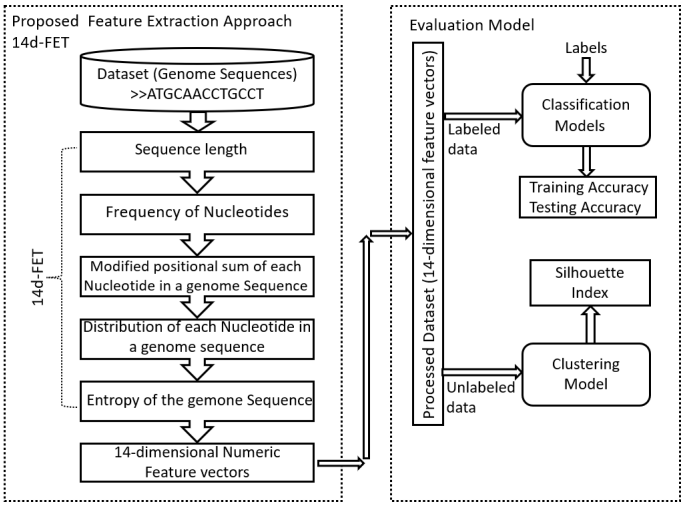

Fig. 1: Proposed 14d-FET Architecture

Algorithm 1 14d-FET

Input: Genome sequence; $x : \{A, T, G, C\}$
Output: Seq_{length}, F_n, MPS_n, Dis_n, Ent_S

1: Initialize the Seq_{length} and F_n with 0, where $n \in \{A, T, G, C\}$
2: Let n is a input character.
3: **for** n in x **do**
$Seg_{length} = Seq_{length} + 1$ (Increase the seq_{len} by 1)
$F_n = F_n + 1$ (Increase the frequency of nucleotide n by 1)
4: **end for**
5: Compute the MPS_n for all nucleotides using Eq. (1).
6: Calculate the Dis_n for all nucleotides using Eq. (2) and (3).
7: Compute the Ent_S using Eq. (4).

a) Sequence length: The sequence length of a genome vary from organism to organism, which provides useful information for categorization and clustering. Hence, in the

proposed approach we extracted sequence length (Seq_{length}) as a feature. The sequence length is the number of nucleotide present in a sequence. For example, a genome sequence is presented in Table I. For this sequence the Seq_{length} will be 9.

TABLE I: Example of genome sequence with a novel power method

	4^0	4^1	4^2	4^3	4^4	4^5	4^6	4^7	4^8
	1	2	3	4	5	6	7	8	9
Seq 1	A	G	T	C	T	A	T	G	C

b) Frequency of nucleotide bases: This type of feature computes the frequency of each nucleotide (F_n). For example, in Table I the number of A, T, G, and C is 2, 3, 2, and 2. Hence, F_A, F_T, F_G, and F_C will have the values 2, 3, 2, and 2.

c) Modified positional sum of nucleotide bases: In this type of feature, we removed the lacunae of 12d-FV [3] approach by fixing the problem of having the same positional sum for a nucleotide n in different sequences by using a new power method, in which we multiply the distances (s_p) by their place values and named these types of features as features based on the modified positional sum. Let us understand the problem of the positional sum of 12d-FV approach [3], if A appears at positions 3 and 4 in one sequence and positions 2 and 5 in another sequence, the positional sum = 7 will be the same for both sequences. Owing to this, the feature based on positional sum cannot correctly differentiate between two distinct genome sequences.

So by using a novel power method, for each nucleotide n, the modified positional sum from the first nucleotide MPS_n is calculated using Eq. (1).

$$MPS_n = \sum_{p=1}^{F_n} s_p * 4^{(s_p-1)}, \tag{1}$$

where s_p is the distance between the p^{th} and first nucleotide. In this technique, we chose four as the base number because there are only four letters in any genome sequence, namely A, T, G, and C. For instance, the nucleotide A appears at locations 1 and 6 in the sequence depicted in Table I, hence the modified positional sum of A is $MPS_A = 1*4^0 + 6*4^5 = 6145$. With this method, the probability of obtaining the same MPS_n for a nucleotide n in two dissimilar sequences is extremely low.

d) Distribution of nucleotide bases: In this type of feature, the distribution of nucleotides is computed by combining features based on frequency (F_n) and features based on modified positional sum (MPS_n). For a nucleotide n, distribution (Dis_n) is calculated by using Eq. (2) and (3).

$$Avgps_n = MPS_n / F_n. \tag{2}$$

$$Dis_n = \sum_{p=1}^{F_n} \frac{(s_p - Avgps_n)^2}{F_n}, \tag{3}$$

where $Avgps_n$ is the average of modified positional sum of a nucleotide n.

e) Entropy: Entropy is used to characterize the amount of randomness in a given sequence. If S is a sequence of nucleotide bases such that $S = N_1, N_2, N_3, ..., N_n$, and let $N_1, N_2, N_3, ..., N_n$ occur in S with probability $P(N_1), P(N_2), P(N_3), ..., P(N_n)$, then the entropy of the sequence (Ent_S) is computed using Eq. (4).

$$Ent_S = -\sum_{i=1}^{n} P(N_i) log P(N_i). \tag{4}$$

After combining all five types of features, the feature vector consists of fourteen features is represented as $< Seq_{length}$, F_A, MPS_A, Dis_A, F_G, MPS_G, Dis_G, F_T, MPS_T, Dis_T, F_C, MPS_C, Dis_C, $Ent_S >$. The feature vector corresponding to the genome sequence shown in Table I will be $<$ 9, 2, 6145, 9418767.25, 2, 131080, 4294836234, 3, 30000, 99900027.66, 2, 590080, 87044766128.5, 1.36 $>$.

The next subsequent section will brief about the experimental findings of the proposed approach on various real life datasets.

IV. Experimental findings

In this study, we collected two types of data, i.e., labeled and unlabeled. In the labeled category, we collected two real-life genome sequence datasets called Bacteria and Fungi from the NCBI repository [7] and one Molecular biology dataset named promoter from the UCI machine learning repository [8]. We have two real-life rice plant SNP datasets in the unlabeled category: SNP-seek rice [9], and MAGIC-rice [10]. We performed classification tasks for the labeled datasets and clustering tasks for the unlabeled dataset. To evaluate the proposed method for labeled data, we have used multiple classifiers named Support Vector Machine (SVM) [11], Decision Tree (DT) [12], Random Forest (RF) [13], Gaussian Naive Bayes (G-NB) [14], Multi-Layer Perceptron (MLP) [15], and K-nearest neighbor classifier (K-NN) [16]. To evaluate the performance on unlabeled data, we have used k-means clustering [17].

This section is divided into four subsections. The first subsection briefs about the datasets used for the experimentation. The second subsection briefs about the performance evaluation measures. The third subsection discusses about the parameter settings for the various parameters. Finally, the fourth subsection explains the experimental findings on the labeled and unlabeled datasets.

A. Dataset details

We used three labeled and two unlabeled datasets in the experimental study presented in Table II. Details of these datasets are discussed subsequently.

The bacteria dataset contains the gene sequences of four bacteria named Bacillus-Subtilis, Aeropyrum-pernix, Aquifex-aeolicus, and Bucchera-sp. The fungi dataset contains the gene sequences of four types fungus named as Cryptococcus, Debaryomyces, Kazachstania africana, and Saccharomyces

TABLE II: Dataset Details

Dataset Name	Count of sequences	Count of classes	Type
Bacteria	340	4	Labeled
Fungi	1051	4	Labeled
Promoter	106	2	Labeled
SNP-seek rice	252	-	Unlabeled
MAGIC-rice	16932	-	Unlabeled

cerevisiae. The promoter dataset contains promoter gene sequences having two types of labels, positive and negative. We converted these symbolic labels to the numerical form of 0 and 1.

The SNP-seek rice data includes information about rice chromosomes 1 through 12. We put all rice chromosomes from ch1 to ch12 into a single file to perform clustering on a large SNP dataset. Details of this data can be found in Mansueto et al. [9]. The MAGIC-rice dataset contains the SNP sequences of rice crops. The MAGIC rice dataset comprises 1,411 samples separated into 12 files (for each chromosome). We put together all chromosome files from 1 to 12 to make the MAGIC-rice dataset. Detailed analysis of the MAGIC-rice dataset is given by Bandillo et al. [10].

B. Evaluation measures

To evaluate the performance of the proposed method, we employed two measures, training accuracy and testing accuracy, for labeled datasets and two metrics, silhouette index and silhouette visualizer, for unlabeled datasets. The details of the metrics are as follows:

a) Accuracy: Accuracy [18] is one of the best parameters to assess the quality of classification models. Accuracy can be defined as the percentage of times our model correctly predicted a given outcome. The accuracy can be computed using Eq. (5). Training accuracy represents the accuracy of the training data, which means the classification model is tested on the same data used in training of the model. Testing or validation accuracy represents the accuracy of testing data. In testing accuracy, the samples used for testing differ from the samples used in training of the model.

$$Accuracy = \frac{Total\ no.\ of\ correct\ predictions}{Total\ no.\ of\ predictions} \quad (5)$$

b) Silhouette index: In unsupervised learning, silhouette index (SI) [19] is widely used as a standard for measurement and analysis. It is based on how similar a data point to the other points in its cluster, which is called cohesion, and to the cluster that is closest to it, which is called separation. The silhouette value ranges from -1 to +1. The SI is calculated by taking the average silhouette coefficient of all data objects.

Let p_i is a i^{th} data point, then its silhouette coefficient S_{p_i} is evaluated by Eq. (6).

$$S_{p_i} = (b_{p_i} - a_{p_i})/(max(b_{p_i}, a_{p_i})), \quad (6)$$

where a_{p_i} is the mean distance of the p_i data point to all other data points within the same cluster, and b_{p_i} is the mean distance of p_i to all other data points in the nearest cluster.

c) Silhouette visualizer: The silhouette visualizer shows which clusters are dense and which are not by displaying the silhouette coefficient per cluster for each sample. It also shows how many clusters have achieved a median SI value.

C. Parameter settings for evaluation models

Various parameters are used in evaluation models to perform the experiments are listed here, and their values are presented in Table III.

TABLE III: Parameter settings for evaluation models

Evaluation Model	**Parameters**	**Parameter Value**
SVM	$Kernel$	rbf
MLP	$Hidden\ layer\ sizes$	$(10, 10, 10)$
MLP	$Activation function$	relu
MLP	$Maximum\ iteration$	$10,000$
DT	$min\ sample\ split$	5
K-NN	K	3
RF	$min\ sample\ split$	5
G-NB	$var\_smoothing$	$1e-9$
k-means	$Maximum\ iteration$	300

D. Experimental Analysis

1) Experimental results on labeled datasets: We run experiments on the labeled datasets listed in Table II and compare the proposed approach to a 12d-FV [3] approach, because both of these methods employ a feature extraction technique for genome sequences. The experiments are performed in a ten-fold cross-validation manner, in which the dataset is divided into ten parts, nine parts are used for training, and one part for testing purposes. The results are quantified in terms of mean training accuracy and mean testing accuracy, which is the mean of training accuracies and testing accuracies of ten folds, respectively.

Experimental results on the bacteria dataset are shown in Table IV. It can be observed that the proposed approach is performing better than 12d-FV [3] in all evaluation models in terms of mean training accuracy and mean testing accuracy. The RF classifier achieves the highest validation accuracy of 92.35%. In comparison to 12d-FV, The proposed approach shows a minimum increment in the mean testing accuracy is of 2.65% when the MLP classifier is used. On the other hand, the proposed approach shows the maximum increment in mean testing accuracy is of 45.1% when the G-NB classifier is used.

TABLE IV: Results on Bacteria Dataset

Evaluation Model	12d-FV		Proposed approach (14d-FET)	
	Mean training accuracy (%)	Mean testing accuracy (%)	Mean training accuracy (%)	Mean testing accuracy (%)
SVM	61.50	60.58	**82.06**	**79.70**
DT	96.92	83.82	**99.11**	**88.23**
RF	98.72	86.47	**99.70**	**92.35**
K-NN	63.87	56.86	**79.54**	**74.50**
MLP	86.63	79.70	**99.11**	**82.35**
G-NB	28.65	20.58	**70.15**	**65.68**

Experimental results on the Fungi dataset are shown in Table V. It can be observed that the proposed approach is performing better than 12d-FV approach [3] in all evaluation models in terms of mean training accuracy and mean testing accuracy. The RF classifier achieves the highest validation accuracy of 93.81%. In comparison to 12d-FV, the proposed approach shows a minimum increment in mean testing accuracy of 0.23% when the MLP classifier is used. On the contrary, the proposed approach shows the maximum increment in mean testing accuracy is of 36.71% when the G-NB is used, which shows the power of proposed approach for enhancing the validation accuracy.

TABLE V: Results on Fungi Dataset

Evaluation Model	12d-FV		Proposed approach (14d-FET)	
	Mean training accuracy (%)	Mean testing accuracy (%)	Mean training accuracy (%)	Mean testing accuracy (%)
SVM	84.62	84.30	**90.71**	**90.20**
DT	98.37	85.44	**98.64**	**92.19**
RF	99.14	89.82	**99.16**	**93.81**
K-NN	94.13	89.81	**95.11**	**91.15**
MLP	94.11	91.81	**96.91**	**92.04**
G-NB	46.32	40.18	**84.37**	**76.89**

Experimental results on the promoter dataset are shown in Table VI. It can be observed that the proposed approach performs better than 12d-FV [3] in all evaluation models in terms of mean testing accuracy. However, in terms of mean training accuracy, the 12d-FV approach is giving higher values than the proposed method for the DT and RF, due to the over-fitting of the model. The proposed method removes this drawback, giving more general models that are not biased towards the training data and produce better results.

TABLE VI: Results on Promoter Dataset

Evaluation Model	12d-FV		Proposed approach (14d-FET)	
	Mean training accuracy (%)	Mean testing accuracy (%)	Mean training accuracy (%)	Mean testing accuracy (%)
SVM	88.05	75.27	**92.13**	**78.27**
DT	**97.59**	65.90	97.06	**72.27**
RF	**98.65**	69.81	98.63	**78.18**
K-NN	78.50	64.18	**85.32**	**72.54**
MLP	99.68	68.63	**99.89**	**73.18**
G-NB	77.57	74.36	**84.06**	**79.09**

2) Experimental results on unlabeled datasets: We performed extensive experiments on the unlabeled datasets listed in Table II for a different number of clusters (k) using k-means and presented the results only few best performing k values. We compared the proposed approach with the 12d-FV [3] approach and presented results in terms of SI and silhouette visualizer.

Table VII displays the results achieved by the proposed 14d-FET on the SNP-seek rice for $k = [2, 10]$. The proposed method yields a greater SI than that of 12d-FV [3] for all k values. It can be seen from the table that $k = 2$ yields the highest SI compared to other k values. So, the silhouette visualizer for 2 clusters, derived from the proposed 14d-FET shown in Fig. 3 demonstrates that all clusters have a higher average SI than the silhouette visualizer presented in Fig. 2 derived from the method of 12d-FV [3].

TABLE VII: SI on k = [2,10] for SNP-seek rice

Number of clusters (k)	**12d-FV**	**Proposed 14d-FET**
2	0.8010	**0.8128**
3	0.7043	0.7589
4	0.6163	0.7187
5	0.5980	0.6839
6	0.5636	0.7369
7	0.5600	0.7067
8	0.5554	0.6832
9	0.5558	0.6654
10	0.5611	0.6300

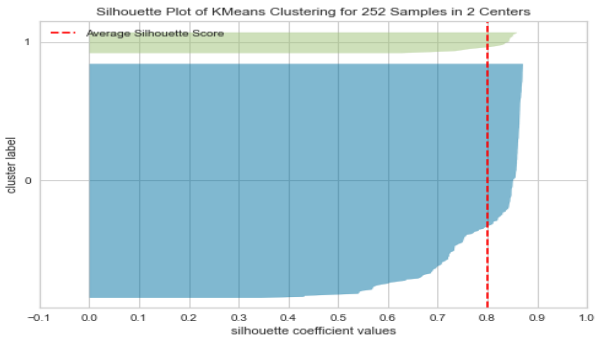

Fig. 2: SNP-seek rice using 12d-FV

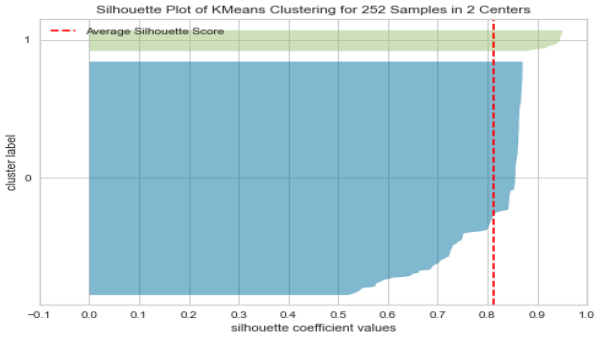

Fig. 3: SNP-seek rice using Proposed 14d-FET

Table VIII displays the results achieved by the proposed 14d-FET on the MAGIC-rice for $k = [2, 10]$. The proposed method yields a greater SI than that of 12d-FV [3] for all k values. It can be seen that proposed 14d-FET yields the highest SI on $k = 3$ compared to other k values. So, the silhouette visualizer for 3 clusters, derived from the proposed 14d-FET shown in Fig. 5 demonstrates that all clusters have a higher average SI than the silhouette visualizer presented in Fig. 4 derived from the 12d-FV method.

V. Conclusion

In this paper, we proposed a novel 14-dimensional feature extraction method, abbreviated as "14d-FET" for genome data made up of the four nucleotide bases A, T, G, and C. Utilizing a novel power method, the proposed 14d-FET tackled the problem of positional sum and distribution features, which can sometimes be the same for dissimilar sequences. In addition, we extracted the most essential features, such as sequence

TABLE VIII: SI on k = [2,10] for MAGIC-rice

Number of clusters (k)	12d-FV	Proposed 14d-FET
2	0.7345	0.7612
3	0.7806	**0.8328**
4	0.6487	0.7178
5	0.6564	0.7067
6	0.6045	0.6614
7	0.4914	0.6609
8	0.4861	0.6919
9	0.4797	0.7003
10	0.4459	0.7244

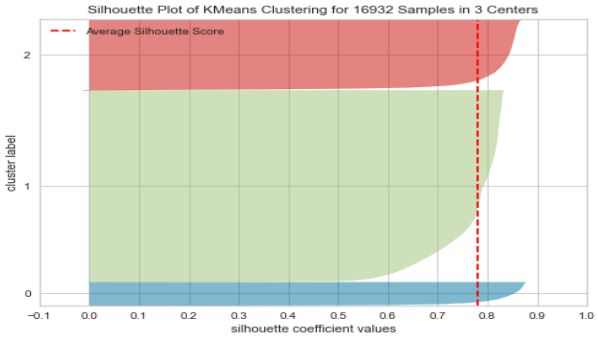

Fig. 4: MAGIC-rice using 12d-FV

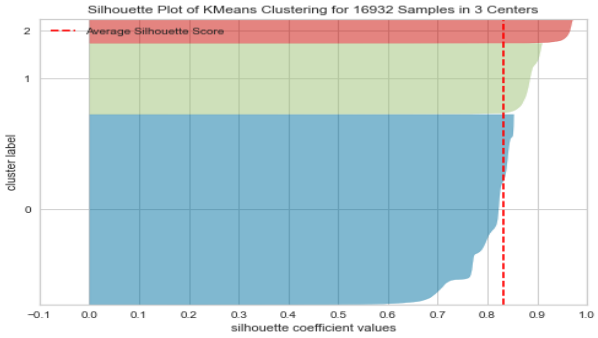

Fig. 5: MAGIC-rice using Proposed 14d-FET

length and entropy, to improve the effectiveness of the proposed feature extraction approach. Moreover, experimental results demonstrate that the proposed method generates the generalized strategy for feature extraction regardless of the evaluation method employed. When the labeled dataset was evaluated on six different classifiers, the proposed method performed better than the existing method. It increased the precision of validation across all labeled datasets. In the case of unlabeled datasets, the proposed method likewise performed well and yielded an enhanced SI compared to existing strategy. Therefore, we can conclude that the proposed method performs well in both supervised and unsupervised learning. In the future, additional feature extraction criteria can be added in proposed feature extraction method to enhance the efficacy of feature extraction method.

ACKNOWLEDGEMENT

This research is funded by The Council of Scientific and Industrial Research (CSIR), Government of India under grant no. 22(0853)/20/EMR-II.

REFERENCES

[1] R. Dwivedi, A. Tiwari, N. Bharill, M. Ratnaparkhe, P. Mogre, P. Gadge, and K. Jagadeesh, "A novel apache spark-based 14-dimensional scalable feature extraction approach for the clustering of genomics data," *The Journal of Supercomputing*, pp. 1–35, 2023.

[2] D. Debroas, J.-F. Humbert, F. Enault, G. Bronner, M. Faubladier, and E. Cornillot, "Metagenomic approach studying the taxonomic and functional diversity of the bacterial community in a mesotrophic lake (lac du bourget–france)," *Environmental microbiology*, vol. 11, no. 9, pp. 2412–2424, 2009.

[3] P. Jha, A. Tiwari, N. Bharill, M. Ratnaparkhe, M. Mounika, and N. Nagendra, "Apache spark based kernelized fuzzy clustering framework for single nucleotide polymorphism sequence analysis," *Computational Biology and Chemistry*, vol. 92, p. 107454, 2021.

[4] M. A. Helaly, S. Rady, and M. M. Aref, *Deep Learning for Taxonomic Classification of Biological Bacterial Sequences*. Cham: Springer International Publishing, 2021, pp. 393–413.

[5] J. Dhaliwal and J. Wagner, "A novel feature extraction method based on highly expressed snps for tissue-specific gene prediction," *Journal of Big Data*, vol. 8, no. 1, pp. 1–13, 2021.

[6] R. P. Bonidia, D. S. Domingues, D. S. Sanches, and A. C. de Carvalho, "Mathfeature: feature extraction package for dna, rna and protein sequences based on mathematical descriptors," *Briefings in bioinformatics*, vol. 23, no. 1, p. bbab434, 2022.

[7] E. W. Sayers, E. E. Bolton, J. R. Brister, K. Canese, J. Chan, D. C. Comeau, C. M. Farrell, M. Feldgarden, A. M. Fine, K. Funk *et al.*, "Database resources of the national center for biotechnology information in 2023," *Nucleic acids research*, vol. 51, no. D1, pp. D29–D38, 2023.

[8] C. Blake, "Uci repository of machine learning databases," *http://www. ics. uci. edu/˜ mlearn/MLRepository. html*, 1998.

[9] L. Mansueto, R. R. Fuentes, F. N. Borja, J. Detras, J. M. Abriol-Santos, D. Chebotarov, M. Sanciangco, K. Palis, D. Copetti, A. Poliakov *et al.*, "Rice snp-seek database update: new snps, indels, and queries," *Nucleic acids research*, vol. 45, no. D1, pp. D1075–D1081, 2017.

[10] N. Bandillo, C. Raghavan, P. A. Muyco, M. A. L. Sevilla, I. T. Lobina, C. J. Dilla-Ermita, C.-W. Tung, S. McCouch, M. Thomson, R. Mauleon *et al.*, "Multi-parent advanced generation inter-cross (magic) populations in rice: progress and potential for genetics research and breeding," *Rice*, vol. 6, no. 1, pp. 1–15, 2013.

[11] R. Dwivedi, R. Kumar, E. Jangam, and V. Kumar, "An ant colony optimization based feature selection for data classification," *Int. J. Recent Technol. Eng*, vol. 7, pp. 35–40, 2019.

[12] Priyanka and D. Kumar, "Decision tree classifier: A detailed survey," *International Journal of Information and Decision Sciences*, vol. 12, no. 3, pp. 246–269, 2020.

[13] T. K. Ho, "Random decision forests," in *Proceedings of 3rd international conference on document analysis and recognition*, vol. 1. IEEE, 1995, pp. 278–282.

[14] K. P. Murphy *et al.*, "Naive bayes classifiers," *University of British Columbia*, vol. 18, no. 60, pp. 1–8, 2006.

[15] H. Taud and J. Mas, "Multilayer perceptron (mlp)," in *Geomatic approaches for modeling land change scenarios*. Springer, 2018, pp. 451–455.

[16] P. Cunningham and S. J. Delany, "K-nearest neighbour classifiers-a tutorial," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–25, 2021.

[17] R. Dwivedi, A. Tiwari, N. Bharill, and M. Ratnaparkhe, "A novel clustering-based hybrid feature selection approach using ant colony optimization," *Arabian Journal for Science and Engineering*, vol. 48, pp. 10 727–10 744, 2023.

[18] H. Bensusan and A. Kalousis, "Estimating the predictive accuracy of a classifier," in *European Conference on Machine Learning*. Springer, 2001, pp. 25–36.

[19] R. Dwivedi, A. Tiwari, N. Bharill, M. Ratnaparkhe, R. Soni, R. Mahbubani, and S. Kumar, "An incremental clustering method based on multiple objectives for dynamic data analysis," *Multimedia Tools and Applications*, pp. 1–21, 2023.