International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Classification of DNA sequences: Analysis of performance of SVM using a novel integral kernel function

Mahesha Y[1], Vishwanath M K[2] , Nagaraju C*[3], Ravindra S[4]

Submitted: 26/01/2024 **Revised**: 04/03/2024 **Accepted**: 12/03/2024

Abstract*:* The DNA sequence classification plays a vital role in bioinformatics to categorize the unknown DNA sequence. In the present article a new approach for DNA classification has been proposed. DNA sequences belong to six different classes are put under the light of present research for classification. A popular classification model SVM has been analysed for exploring the insights of it. Two kernel functions namely Radial Basis and Polynomial are kept under deep analysis for integrating these kernels. These two kernel functions are successfully integrated. An experiment has been carried out by integrating Radial Basis Function with Polynomial kernel. The performance has been measured using metrics such as Precision, Recall, F1-score, and Accuracy. The performance of the methods adopted in the present research has also been shown using Precision-Recall curve and ROC curve. The individual accuracy achieved by Radial Basis Function and Polynomial function is 80.6% and 82.7% respectively. The proposed novel model has achieved accuracy of 98.4%. The result clearly shows that the proposed integral kernel has outperformed the Radial Basis and Polynomial functions.

Keywords: *DNA, K-mer, Machine learning, RBF, Polynomial, Integral Kernel*

1. Introduction

DNA is a polymer comprised of two polynucleotide chains that are arranged in a double helix. The organisms have genetic data in DNA which is responsible for growth and other operations of the organism. The strands of DNA composed of nucleotides named as as polynucleotides. The nucleotide is composed of phospohate, sugar and nucleobases. The nucleobases are cytosine [C], guanine [G], adenine [A] and thymine [T]. The nucleotides are joined together to form backbone of DNA. The strands are connected by hydrogen bonds according to pairing principle [1, 2]. A standard laboratory procedure for identifying the precise arrangement of bases, or nucleotides, in a DNA molecule is named as "DNA sequencing." The base sequences contain the essential information that cells need for their function. The bases are represented as A, T, C, and G. The DNA sequencing is necessary to comprehend how gene and other parts of the genome function [3].

DNA sequence categorization is the process of determining the extent to which an unidentified sequence S belongs to an already-existing class C. Classification is a crucial machine learning research activity [4, 5]. Its objective is to create a classification method using the training data set to predict how future, unlabeled items will be classified. The categorization of known genes as a certain type of data is a problem that frequently arises in knowledge discovery [6, 7, 8]. By identifying the kind of DNA sequence dependent around how identical its organization or functioning is with that of other sequences, sequence classification aids in finding the genetic traits in nucleotide sequences. It also predicts the function of each sequence and how they relate to one another.

The SVM method predicts whether a new instance falls into the same category or a different one by representing the data in the space of n dimensions [9, 10]. The SVM works with kernel system for classification. The most important kernels with which SVM works are Gaussian kernel, Radial Basis Function, Sigmoid kernel and Polynomial kernel. Among these kernel functions two kernels namely Radial Basis Function and Polynomial have been taken for analysis. Classification problems can be effectively handled by the potent machine learning method known as Radial Basis Function Support Vector Machine (RBF SVM). This non-parametric model performs effectively when dealing with high-dimensional, non-linear data. The way that RBF SVM operates is by projecting the input data onto a higher-dimensional feature space, which allows a hyperplane to divide the classes. The algorithm calculates the similarity between pairs of data points in the feature space using a kernel function, such as the Radial Basis Function. Another well known kernel function is polynomial kernel function. Considering the polynomial functions of the original features, the polynomial kernel transfers the input data into a higher-dimensional feature space.

[1] *Mysuru Royal Institute of Technology, Mandya, India*
[2] *The National Institute of Engineering, Mysore, India*
[3*] *The National Institute of Engineering, Mysore, India*
[4]*City Engineering College, Bangalore, India*
** Corresponding Author Email: nagaraj@nie.ac.in*

2. Literature review

Circular graphs were presented as a model of DNA sequence for classification [11]. Here, a query DNA was classified by matching it with words stored in the database. This method has achieved 94% of accuracy. A method was presented for categorizing the DNA sequences based on vector space [12]. In this method, the author has used Principal Component Analysis (PCA) for DNA classification. This approach was able achieve accuracy of 96% in the classification of extrons and introns. A classification model was proposed to identify DNA sequence which includes *E. Coli sequences* [13]. This model is based on neural network and the author has claimed that the proposed model outperformed other existing prediction systems. They used four coding approaches for encoding the DNA sequences and four neural networks was trained using the encoded DNA sequence. The results of four neural networks were integrated using logarithm function. The primary flaw in the neural network architecture is the problem in acquiring the optimal neural network parameters.

To categorize the DNA sequences of *E.Coli* promoters a classification model was proposed which adopts expectation-maximization [14]. In this approach, a new DL model was proposed. Here, author has concluded that main weakness of the DL method is need of data for training the system. He used enhanced expectation-maximization method for finding 35 and 10 adhering sites in the *E.coli* promoter sequence. It is no more believed that there will be a consistent arrangement of spacer lengths across binding sites. The probability model of the lengths was derived by him. In line with the details given in the sequence, he selected and applied orthogonal computational model to represent the features. Subsequently, the features are sent to neural network for sequence recognition. With a number of datasets, this strategy worked effectively.

Hidden Markov model called VOGUE was proposed [15] which use changeable sequence mining method to find interesting patterns with varied durations and distances across the elements. The accuracy of the model is greater than the traditional hidden markov model. The model's ability to generalise is affected by the fact that the frequent patterns of the components in the sequence do not included in the model.

In recent times, the convolutional neural networks (CNN) have gained popularity as a deep learning model [16]. The CNN can be employed to examine data and derive semantic information [17, 18, 19]. A model which employed DNA sequences as textual information and developed a unique approach for categorizing DNA sequences using CNN [20]. This technique utilizes a one-stop vector to express the sequence as the model's input. As a outcome, it records the data for every nucleotide in the basic location sequence. Model has been assessed using twelve data sets of sequences and the result has shown that model has significantly enhanced on every dataset. The model has achieved cross validation accuracy of 85.41%. The continued progression of deep learning produced innovative strategies for DNA sequence analysis. In supervised machine learning classifiers, feature extraction is an important pre-requisite. Two unique DL models presented and their performance was evaluated using five datasets [21]. The experiment has shown that neural networks have the ability to extract useful features from the input data.

In the present article, Support Vector Machine (SVM) has been explored for classification of DNA sequences. Two kernel functions namely Radial Basis Function (RBF) and Polynomial function have been considered for developing a new kernel function for SVM. A new kernel is obtained by integrating RBF with Polynomial function as explained in the next section.

3. Materials and Methods

The data set of DNA sequence has been used been in the experiment. The data set has 4380 DNA sequences belong to seven classes. The classes are labelled from class 0 to class 6. There are 531 sequences belongs to class 0, 534 sequences belongs to class 1, 349 sequences belong to class 2, 672 sequences belong to class 3, 711 sequences belong to class 4, 240 sequences belong to class 5 and 1343 sequences belong to class 6.

The present article explores the performance of novel kernel in SVM for classification of DNA sequences. The kernel functions namely RBF kernel [22], Polynomial kernel [23] and Integral kernel have been experimented to spread light on their ability to perform DNA classification. The DNA classification needs few steps as shown in Fig. 1.

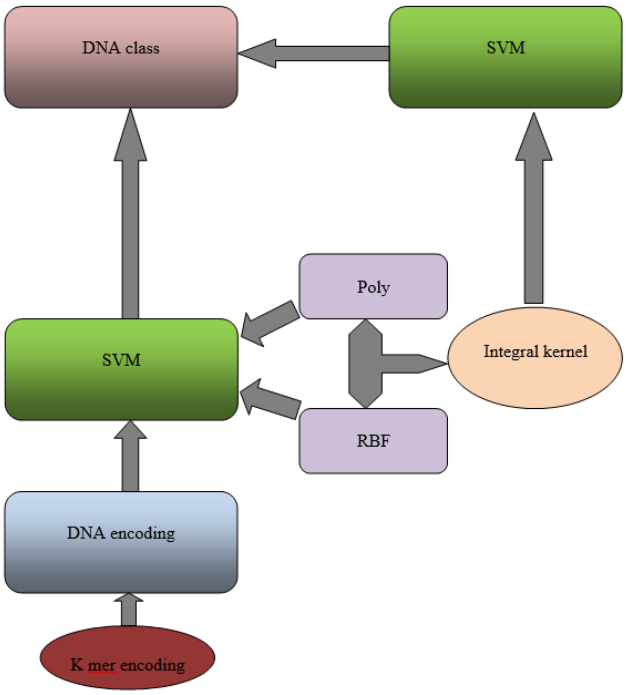


Fig. 1. The phases of DNA classification

3.1 DNA encoding

The DNA sequence must be transformed into specific value to make it suitable for providing it as input to the classifiers. Sequential encoding, one-hot encoding, and k-mer encoding are the methods of encoding can be used for DNA sequencing [24]. In the present experiment, k-mer encoding has been used.

3.1.1 k-mer encoding

K-mers are k-length substrings of a genetic pattern that are used in bioinformatics [25]. K-mers, which are made up of nucleotides (such as A, T, G, and C) and are mostly utilised in the field of computational genomics and sequence analysis, are used to construct DNA sequences, enhance differential gene manifestation, discover organisms in metagenomic materials, and develop weakened vaccines. Typically, the phrase "k-mer" corresponds to each k-mer fragment in a string.

3.2 Integral Kernel

The Radial basis function given in Eq. (1) has been integrated with polynomial function given in Eq. (2) to obtain a novel integral function shown in Eq. (3).

$$e^{\frac{-1}{2}\|x-x'\|^2} = \sum_{n=0}^{\infty} \frac{x^n}{n!}$$

$$= \sum -x^T x' - \frac{1}{2}\|x\|^2 - \frac{1}{2}\|x'\|^2$$

$$= \sum_{j=0}^{\infty} (x^T x')^j \sum_{n=0}^{\infty} \frac{\frac{-1}{2}\|x\|^2}{n!} \sum_{n=0}^{\infty} \frac{\frac{-1}{2}\|x'\|^2}{n!} \quad (1)$$

$$= < z(x), z(x') >$$

Where,

$$z(x) = \sum_{n=0}^{\infty} \frac{\frac{-1}{2}\|x\|^2}{n!}$$

Polynomial function,

$$k(x^i, x^j) = (x^{i^T} x^j + 1)^d$$

$$= (\sum x_i y_i + c)^2$$

$$= \sum (x_i)^2 (y_i)^2 + \sum_{i=2}^{n} \sum_{j=1}^{i-1} (\sqrt{x} x_i x_j)(\sqrt{2} y_i y_j) + \sum (\sqrt{2c} x_i) * (\sqrt{2c} y_i) + c^2 \quad (2)$$

Following function given in Eq. (3) has been used as a kernel in SVM for DNA classification,

$$k(z(x)) = (< z(x), z(x') >^T < z(x), z(x') > +1)^2$$

$$= (\sum < z(x), z(x') >^T < z(x), z(x') > y_i + c)^2$$

$$= \sum (\sum_{0}^{\infty} \frac{\frac{-1}{2}\|x\|^2}{n!})^2 + \sum_{2}^{n} \sum_{j=1}^{i-1} \sqrt{2} \sum_{0}^{\infty} \frac{\frac{-1}{2}\|x_i\|^2}{n!} \sum_{0}^{\infty} \frac{\frac{-1}{2}\|x_j\|^2}{n!} \sqrt{2} y_i y_j + \sum \sqrt{2c} \sum_{0}^{\infty} \frac{\frac{-1}{2}\|x_i\|^2}{n!} \sqrt{2c} y_i + c^2 \quad (3)$$

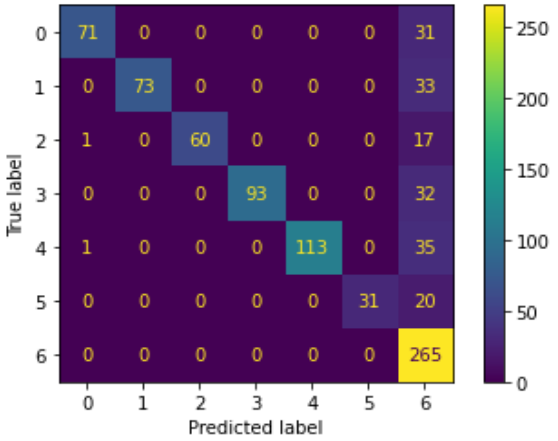
4. Results

As a preprocessing step the DNA sequences have been encoded. For this k-mer encoding was used. The value of k is taken 6 and hence it is called 6-mer encoding. The encoding of the training sequences into k-mer words of length six is done as shown in Fig.4.

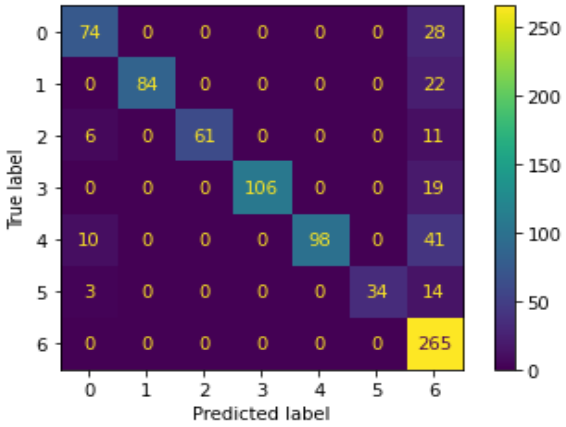
```
0       [atgccc, tgcccc, gcccca, ccccaa, cccaac, ccaac...
1       [atgaac, tgaacg, gaacga, aacgaa, acgaaa, cgaaa...
2       [atgtgt, tgtgtg, gtgtgg, tgtggc, gtggca, tggca...
3       [atgtgt, tgtgtg, gtgtgg, tgtggc, gtggca, tggca...
4       [atgcaa, tgcaac, gcaaca, caacag, aacagc, acagc...
```

Fig. 2: K-mer encoding of training data

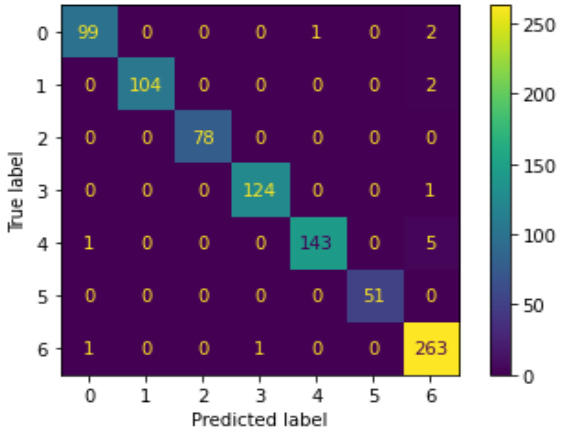
The dataset is divided into training and testing set in the ratio 80:20. The encoded training set is provided as input to the SVM with different kernels. The result of each kernel is discussed below. The confusion matrix for RBF, Polynomial and Novel Integral Kernel has been presented in Fig. 3. The performance of the kernel functions has been evaluated using metrics such as Precision, Recall, F1-score and Accuracy and the results have been tabulated in Table 1.



(a) RBF kernel



(b) Polynomial kernel



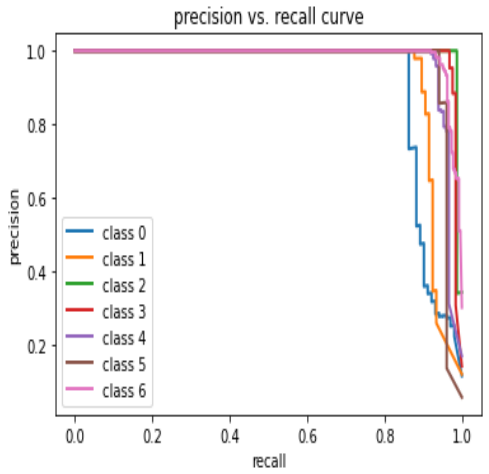
(c) Integral kernel

Fig. 3: Confusion matrix for human DNA sequence classification using (a) RBF Kernel (b) Polynomial Kernel (c) Integral Kernel

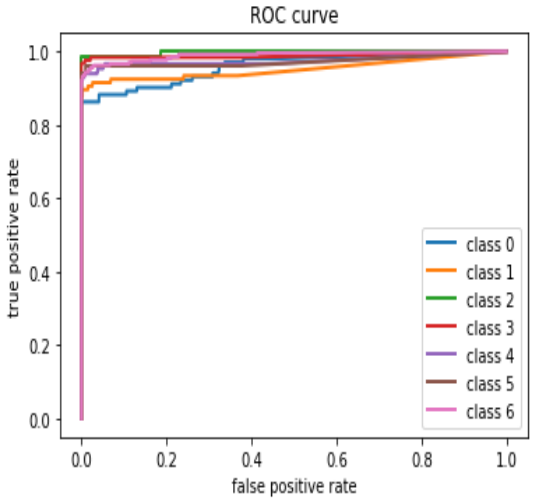
Table 1. Performance of classifiers

Classifier	Precision	Recall	Accuracy	F1-score
RBF	87.9	80.6	80.6	81.3
Polynomial	87.4	82.4	82.4	82.7
Integral kernel (RBF+Polynomial)	93.7	92.3	98.4	92.4

From the results generated as presented in Table 1 it is found that developed kernel is producing desirable results. The PR and ROC curves have been plotted for RBF, Polynomial and Integral kernels and the same has been presented in Fig. 4 to Fig. 6. The precision, recall, accuracy and F1 score are 87.9, 80.6, 80.6 and 81.3 respectively for RBF, 87.4, 82.4, 82.4 and 82.7 for Polynomial, and 93.7, 92.3, 98.4 and 92.4 for integral kernel.



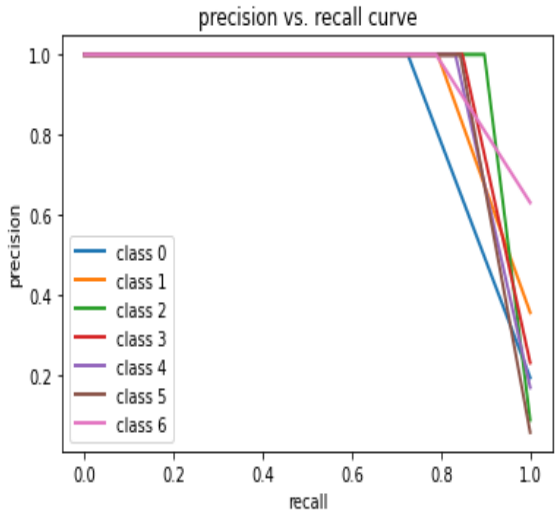
(a)



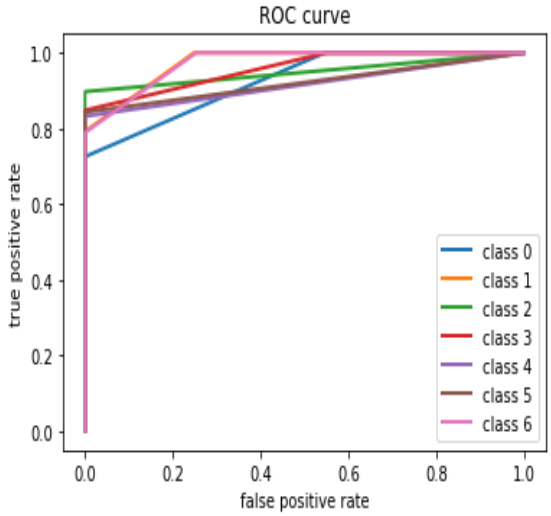
(b)

Fig. 4: (a) PR curve (b) ROC Curve for novel integral kernel

function

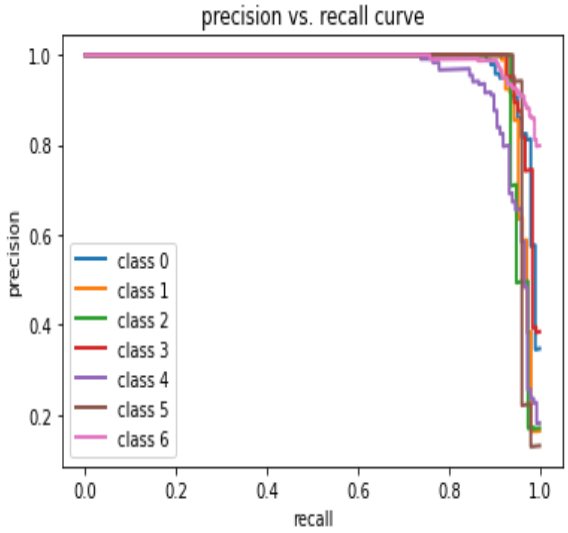


(a)

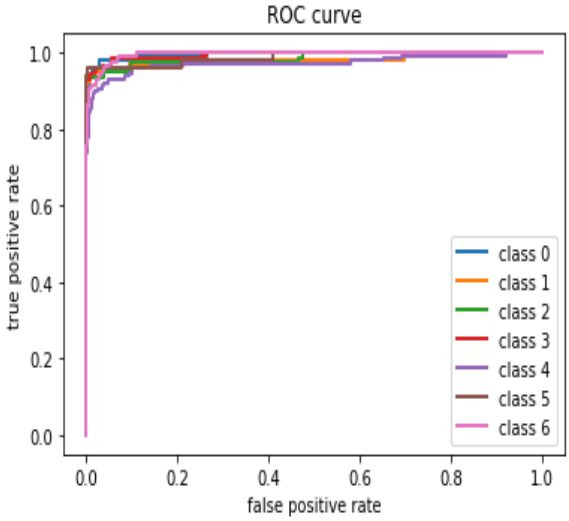


(b)

Fig. 5: (a) PR curve (b) ROC Curve for Polynomial



(a)



(b)

Fig. 6: (a) PR curve (b) ROC Curve for RBF

The AUC values of the kernel functions are tabulated in Table 2.

Table 2: AUC of classifiers

Kernel	AUC							
	C1	C1	C2	C3	C4	C5	C6	Avg.
RBF	0.95	0.95	0.99	0.98	0.97	0.97	0.99	0.97
Poly	0.92	0.97	0.94	0.95	0.91	0.92	0.97	0.94
Novel Kernel	0.99	0.98	0.98	0.99	0.97	0.98	0.99	0.98

The results have shown that all the classifiers are suitable for classification of DNA sequences. The accuracy achieved by SVM with RBF kernel, Polynomial kernel and Novel Integral Kernel is presented in Table 3.

Table 3: Comparison of novel kernel with RBF and Polynomial kernels

Kernal	Accuracy
RBF	80.6
Polynomial	82.4
Novel Integral Kernel	98.4

5. Conclusion

The DNA sequences are classified using SVM classifier using three different kernels. The existing kernels namely RBF and Polynomial are used to test the performance of SVM. The RBF kernel and polynomial kernel have achieved accuracy of 80.6 and 82.4 respectively. These two kernels are analyzed for integration to achieve better accuracy. The RBF kernel is integrated with Polynomial kernel to get a new kernel for SVM function. The resultant kernel has achieved accuracy of 98.4%. The ROC curve has also been drawn and AUC values have been analyzed. The AUC values of RBF, Polynomial and Integral kernels are 0.97, 0.94 and 0.98 respectively. The overall analysis shows that there exists a space to integrate kernel functions to get better accuracy in classification problems.

References

[1] Alberts B, Johnson A, Lewis J, et al. Molecular Biology of the Cell,4th edition, New York: Garland Science, 2002, The Structure and Function of DNA.

[2] Ghannam JY, Wang J, Jan A. Biochemistry, DNA Structure, In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2022.

[3] Brown TA. Genomes. 2nd edition. Oxford: Wiley-Liss; 2002. Chapter 7, Understanding a Genome Sequence

[4] Mahesha Y, Nagaraju C , A literature review on analysis of palm patterns to detect congenital heart diseases, Biomed Eng Appl Basis Commun 32, 2020, https://doi.org/10.4015/S101623722050012

[5] Y. Mahesha and C. Nagaraju, "Automating the Identification and Evaluation of the Position of Axial Triradius on Palm Print: An Approach to Early Detection of Congenital Heart Diseases", Biomedical Engineering: Applications Basis and Communications, vol. 33, no. 2, pp. 2150021, Apr. 2021.

[6] Mahesha, Y., Nagaraju, C. (2023). Analysis of Axial Triradius to Detect Congenital Heart Diseases. In: Ranganathan, G., Bestak, R., Fernando, X. (eds) Pervasive Computing and Social Networking. Lecture Notes in Networks and Systems, vol 475. Springer, Singapore. https://doi.org/10.1007/978-981-19-2840- 6_2

[7] Mahesha Y and Nagaraju C, "Principal Component Analysis and Local Binary Patterns: A comparative study using different databases", International Research Journal of Modernization in Engineering Technology and Sciences", 5(11), 2023. https://www.doi.org/10.56726/IRJMETS46856

[8] Yang A, Zhang W, Wang J, Yang K, Han Y and Zhang L (2020) Review on the Application of achine Learning Algorithms in the Sequence Data Mining of DNA. Front. Bioeng. Biotechnol. 8:1032

[9] Warjurkar, S. V. ., & Ridhorkar, S. . (2024). Maximizing Precision in Early Prognosis using SVM-ACO Classifier and Hybrid Optimization Techniques in MRI Brain Tumor Segmentation with Integration of Multi-Modal Imaging Data. International Journal of Intelligent Systems and Applications in Engineering, 12(10s), 389–401.

[10] Kumar, A. ., Gaur, N. ., & Nanthaamornphong, A. . (2024). Intelligent Signal Identification of NOMA Signal with 256-QAM Modulation Using SVM Algorithm. International Journal of Intelligent Systems and Applications in Engineering, 12(13s), 257–264.

[11] Levy, S., Stormo, G.D. (1997). DNA sequence classification using DAWGs. In: Mycielski,J., zenberg, G., Salomaa, A. (eds) Structures in Logic and Computer Science. Lecture Notes in Computer Science, vol 1261. Springer, Berlin, Heidelberg.

[12] H.-M. Müller, S.E. Koonin, "Vector space classification of DNA sequences", Journal of Theoretical Biology, 223(2), 2003, pp. 161-169.

[13] Ranawana, R., Palade, V. A neural network based multi-classifier system for gene identification in DNA sequences. Neural Comput & Applic 14, 122–131 (2005).

[14] Samia M. Abd –Alhalem, El-Sayed M. El-Rabaie, Naglaa. F. Soliman, Salah Eldin S. E. Abdulrahman, Nabil A. Ismail and Fathi E. Abd El-samie, DNA Sequence Classification with Deep Learning: A Survey, 2020

[15] Zaki MJ, Carothers CD and Szymanski BK, " VOGUE: A Variable Order Hidden Markov Model with Duration Based on Frequent Sequence Mining", ACM Transactions on Knowledge discovery from Data, 4(1), 2010.

[16] Y. Mahesha and C. Nagaraju, "Spotting congenital heart diseases using palm print based on faster R-CNN and spatial method, International Journal

of Medical Engineering and Informatics 2024 16:1, 56-70. https://doi.org/10.1504/IJMEI.2024.135685.

[17] Y. Mahesha and C. Nagaraju, “Machine learning approach to detect congenital heart diseases using palmar dermatoglyphics”, International Journal of Medical Engineering and Informatics 2023 15:4, 336- 351, https://doi.org/10.1504/IJMEI.2023.132575

[18] Mahesha, Y. (2023). Identification of Brain Tumor Images Using a Novel Machine Learning Model. In: Ranganathan, G., Papakostas, G.A., Rocha, Á. (eds) Inventive Communication and Computational Technologies. ICICCT 2023. Lecture Notes in Networks and Systems, vol 757. Springer, Singapore. https://doi.org/10.1007/978-981-99-5166-6_30

[19] M. Y and N. C, "Machine Learning Approach to Detect Congenital Heart Diseases using Angle at Axial Triradius," 2021 IEEE Mysore Sub Section International Conference (MysuruCon), Hassan, India, 2021, pp. 220-226, doi: 10.1109/MysuruCon52639.2021.9641585.

[20] Le NQK, Yapp EKY, Nagasundaram N, Yeh HY. Classifying Promoters by Interpreting the Hidden Information of DNA Sequences via Deep Learning and Combination of Continuous FastText N-Grams. Front Bioeng Biotechnol. 2019 Nov 5;7:305.

[21] Di Gangi M, Lo Bosco G, Rizzo R. Deep learning architectures for prediction of nucleosome positioning from sequences data. BMC Bioinformatics. 2018 Nov 20;19(Suppl 14):418

[22] Hui Cao, Takashi Naito and Yoshiki Ninomiya, "Approximate RBF Kernel SVM and Its Applications in Pedestrian Classifcation", Low-Power High-Speed ADCs for Nanometer CMOS Integration, 2008.

[23] Rikard Vinge and Tomas Mckelvey, “Understanding Support Vector Machine with Polynomial Kernels”, 2019 27th European Signal Processing Conference (EUSIPCO), A Coruna, Spain, 2019, pp. 1-5

[24] Periwal N, Sharma P, Arora P, Pandey S, Kaur B and Sood V, “A novel binary k-mer approach for classification of coding and non-coding RNAs across diverse species”, Biochimie, 199, pp. 112-122, 2022.

[25] Orozco-Arias S, Candamil-Cortés MS, Jaimes PA, Piña JS, Tabares-Soto R, Guyot R, Isaza G. K-mer-based machine learning method to classify LTR-retrotransposons in plant genomes. PeerJ. 2021 May 19;9:e11456