Cell Biochemistry and Biophysics
https://doi.org/10.1007/s12013-024-01336-3

ORIGINAL PAPER

A Kernelized Classification Approach for Cancer Recognition Using Markovian Analysis of DNA Structure Patterns as Feature Mining

Vijay Kalal[1] · Brajesh Kumar Jha[1]

Accepted: 22 May 2024

Abstract
Nucleotide-based molecules called DNA and RNA are essential for several biological processes that affect both normal and cancerous cells. They contain the critical genetic material needed for normal cell growth and functioning. The DNA structure patterns that make up the genetic code affect cells' growth, behavior, and control. Different DNA structure patterns indicate different physiological effects in the cell. Knowledge of these patterns is necessary to identify the molecular origins of cancer and other disorders. Analyzing these patterns can help in the early detection of diseases, which is essential for the effectiveness of cancer research and therapy. The novelty of this study is to examine the patterns of dinucleotide structure in many genomic regions, including the non-coding region sequence (N-CDS), coding region sequence (CDS), and whole raw DNA sequence (W.R. sequence). It provides an in-depth discussion of dinucleotide patterns related to these diverse genetic environments and contains malignant and non-malignant DNA sequences. The Markovian modeling that predicts dinucleotide probabilities also reduces feature complexity and minimizes computational costs compared to the approaches of Kernelized Logistic Regression (KLR) and Support Vector Machine (SVM). This technique is effectively evaluated in essential case studies, as indicated by accuracy metrics and 10-fold cross-validation. The classifier and feature reduction, which are generated by Markovian probability, operate well together and can help predict cancer. Our findings successfully distinguish DNA sequences related to cancer from those diagnostics of non-cancerous diseases by analyzing the W.R. DNA sequence as well as its CDS and N-CDS regions.

Keywords Cancer and non-cancer · Dinucleotide analysis · KLR and SVM · Nucleotide sequences · Markov chain model

Introduction

The nucleus is the core of the cell, along with numerous other essential parts. A human cell consists of 23 pairs of chromosomes, and these chromosomes contain a variety of genes. DNA, or "deoxyribonucleic acid", is a huge double-helix molecule that makes up the genes. It stores genetic information in DNA, which is the basic biological macromolecule. The sugar phosphate and nitrogenous bases (also known nucleotide pairs) that make up the "rungs" and this nitrogenous basis of the ladder are adenine (A), thymine (T), cytosine (C), and guanine (G), as shown in Fig. 1. The complementary nature of these nucleotides is noteworthy, as A combines a pair with T, and C combines a pair with G [1–6].

Cancer is a disease that affects people all over the world and is caused by abnormalities in cells. It may be distinguished from normal cell behavior by the complex changes that occur inside the cells. The complexity of the disease is apparent given the over 100 forms that have been found, which include skin, lung, prostate, ovarian, breast, and skin cancers. The deadly nature of cancer is mostly due to the instability of genes such as cell cycle regulators, tumor suppressors, and proto-oncogenes. Conventional therapies such as chemotherapy are expensive and associated with a great deal of side effects, but their effectiveness is restricted. This highlights the pressing need for fewer surgeries and more robust therapies to address this leading cause of death in affluent nations [7–11]. Genetic disorders can arise from "mutations" or variations in the nucleotide sequence. These changes to the nucleotide

✉ Brajesh Kumar Jha
brajeshjha2881@gmail.com
Vijay Kalal
vijay.kphd21@sot.pdpu.ac.in

[1] Department of Mathematics, School of Technology, Pandit Deendayal Energy University, Raysan, Gandhinagar, Gujarat 382007, India

Published online: 07 June 2024

sequence may affect the overall gene sequence. Furthermore, certain genetic diseases may not just result from nucleotide changes; inherited features, environmental variables, and epigenetic modifications can all add to the complexity of genetic disorders [12, 13].

The raw DNA sequence consists of nucleotides that are added to the 3′ end of the building helix, as DNA is always synthesized in a 5′-to-3′ orientation. There are two regions of the raw DNA sequence: the non-coding region (denoted as N-CDS) and the coding region (denoted as CDS) which is shown in Fig. 2. The nucleotide sequence that codes for proteins is known as the Coding Region Sequence (CDS), whereas the nucleotide sequence that does not code for proteins is known as the Non-Coding Region Sequence (N-CDS) [14]. The whole raw sequence of DNA (denoted as W.R Sequence), includes the CDS and N-CDS parts, as shown in Fig. 2.

Data mining has been used extensively to study DNA sequences, both coding and non-coding, in the context of cancer. These studies typically require the analysis of specific genes, including mutations that are cancerous and non-cancerous. Over last decade, a lot of scientific work has focused on the analysis of DNA sequences to find unique biological patterns. These patterns involve identifying the locations of genes and employing DNA coding sequence regions to distinguish between cancerous and noncancerous DNA sequences [4]. DNA sequence study is currently included in big data analysis due to the exponential expansion of DNA sequences. The rapid developments in sequencing technology cause the number of DNA sequence data [3, 4, 6]. The study uses a variety of computational methods and signal processing to extract characteristics [15, 16]. Using the electrical stimulation of genomic sequence subunits as a basis for segmenting categorization procedures is a unique method for classifying sequence-type data. Furthermore, Roy et al. effectively implemented this idea on many datasets [17, 18]. Several strategies for improving performance were put forth, and the integration of earlier concepts was investigated to signal better processing approaches employing computational techniques. Both Das and Barman as well as Roy & Barman looked at the corresponding amino acid analysis of genomic sequences [19, 20]. Cancerous and non-cancerous DNA sequences have significantly varying nucleotide lengths in cancer classification. Data pre-processing is an important stage in classification challenges, such as utilizing machine learning models for cancer diagnosis from DNA sequences. ‘iACP-GAEnsC’ is a sophisticated model for anticancer peptide identification obtained by Shahid Akbar et al. and it combines three distinct feature representation techniques for protein sequences with evolutionary intelligent genetic algorithms [8]. Later, based on their use of FastText embeddings and a deep neural network to distinguish ACPs, Shahid Akbar et al. presented the cACP-DeepGram model for medication creation and scientific study [9, 10]. subsequently following Shahid Akbar et al.‘s analysis of peptide encoding techniques, with a particular emphasis on KSAAP’s efficiency. To improve the performance of their model, they test learning hypotheses [11]. A multitude of other scholars also proposed pre-processing techniques. Wei Huang et al. used max-min normalization in their study’s pre-processing data to evaluate the

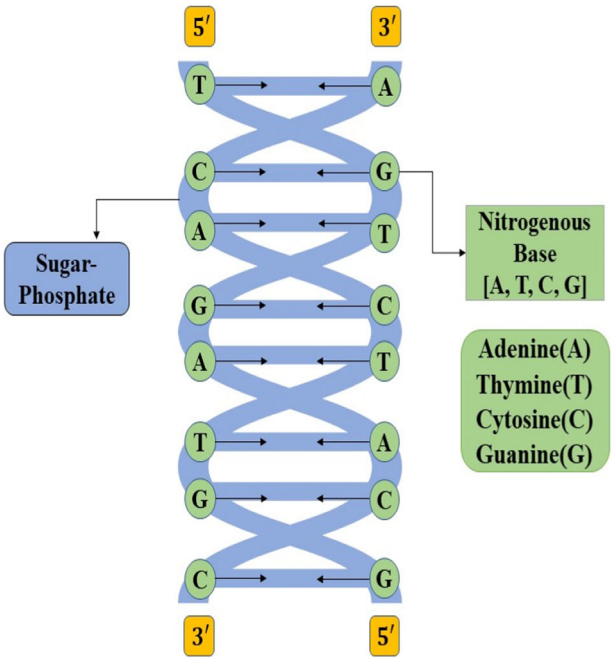

Fig. 1 Nucleotide Sequence in the form Ladder Shape “DNA”

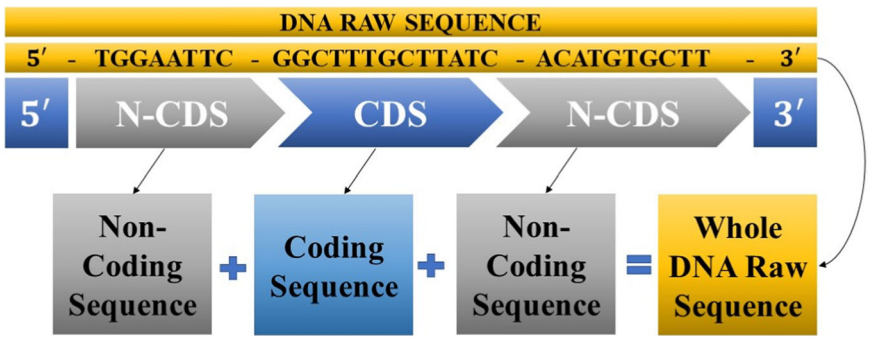

Fig. 2 Raw DNA Sequence with CDS and Non-CDS

similarity of biological sequences [2]. An important problem is the categorization of nucleotide sequences as cancerous or non-cancerous based on nucleotide pairs found in the DNA sequences of certain genes. According to the studies by K. Kourou et al., M. Margaliot et al., N. SenthilVelMurugan et al., A. A. T. Fernandes et al., L. Liu et al., the comprehension of feature extraction development in a large number of datasets and classifying the binary output value can be improved by using machine learning (ML) and data mining techniques, but these techniques require a sufficient degree of validation before they can be used in routine real-world application practice [21–30]. Maverick Lim Kai Rong et al. investigated the nucleotide mutation rate using the Kimura-2 parameter model, in the time series and spatial domains of the SARS-CoV-2 genome sequence as a stochastic process [31]. By using logistic regression-based techniques for image reconstruction in EIT and UST, Tomasz Rymarczyk et al. aimed to improve industrial tomography [32]. Amin Khodaei et al. developed a feature extraction method for classifying and detecting cancerous DNA sequences based on the Markov chain to overcome this difficulty. To address this problem, researchers developed a pattern recognition model that uses signal processing and support vector machines to discriminate between DNA sequences that are cancerous and non-cancerous [4, 5]. Applications such as DNA sequence chain assembly are used to identify genes and estimate the locations of protein-coding regions [33]. According to the reviewed study, the comparative classification focuses only on coding regions within DNA sequences. However, to advance the scope of this study, non-coding DNA sequences must be included, as well as whole raw sequences containing both coding (CDS) and non-coding (N-CDS) regions.

We analyze a novel technique for feature extraction and selection based on the first-order Markov chain of nucleotides. This approach modifies adjacent nucleotide probabilities, with a particular emphasis on a dinucleotide probability distribution analysis in DNA sequence. Our research clarifies the application of dinucleotide probability as a feature in large-scale DNA sequence datasets for the classification of cancer. The main objective of this work is to employ this novel approach to analyze by dividing nucleotide DNA sequences into groups that correspond to protein regions, non-protein regions, and both combined regions. This includes information on DNA sequences linked to cancer and non-cancer conditions. These sequences provide essential information for comparing and predicting both cancerous and non-cancerous DNA sequences via Kernel Logistic Regression (KLR) and Support Vector Machines (SVM). The remaining sections of this paper are organized as follows: The useful tools, fundamental concepts, and algorithms used in the study are described in depth in "Materials and Methods". A thorough explanation of the approach and a presentation of the results analysis are provided in "Result and Discussion" and "Discussion".

Materials and Methods

This work describes the modeling and analysis of DNA nucleotide sequences, as well as the computational and statistical mapping of whole raw DNA sequences, non-coding sequences (N-CDS), and coding sequences (CDS). A classifier that uses the Kernel logistic regression (KLR) and support vector machine (SVM) is combined with a feature extraction methodology based on the first-order Markov chain of nucleotides in this hybrid approach. The method uses the Markov chain of nucleotides, more precisely the dinucleotide analysis, for feature selection and extraction. Moreover, KLR and SVM are used in the comparative analysis to classify the samples according to the defined features. Additionally, a pattern recognition technique for distinguishing between cancerous and non-cancerous genes is proposed.

The suggested algorithm's basic phases are shown in Fig. 3 as a flowchart. This method extracts features using the first-order Markov chain of nucleotides. Case studies are classified using a non-linear kernel function method after an efficient feature selection strategy has been used. Standard criteria are used for evaluation, including the major metrics TP, TN, FP, and FN, as well as supplementary metrics including F1-Score, accuracy, specificity, recall, and precision. 10-fold cross-validation is used to improve the suggested model evaluation method. The approaches and procedures will be thoroughly explained in the parts that follow.

Data Compilation (Case Studies)

GenBank, a database maintained by the NCBI, offered sample data for analysis and comparisons [34]. 338 cases were used to categorize data and evaluate the accuracy of the suggested approach, including CDS, N-CDS, and whole raw sequences. In the selected samples, there is about an equal distribution of cancerous and non-cancerous instances. In particular, genes connected to prostate, colon, and breast cancer are linked to the selected DNA nucleotide sequence samples. Additionally, these genes are chosen without considering the location of the human chromosome. The outcomes of these analyses, which were conducted with 1111 samples, will then be provided in the discussion part that follows. Table 1 provides quantitative information from case studies from the literature [4, 5, 15–20].

Pattern Recognition Via Nucleotide Sequence Mining

Diagnosing specific data and classifying it into two or more groups is an important step in pattern recognition. The process of identifying patterns is based on a criterion for discriminating that is obtained from the similarity between the characteristics that have been extracted.

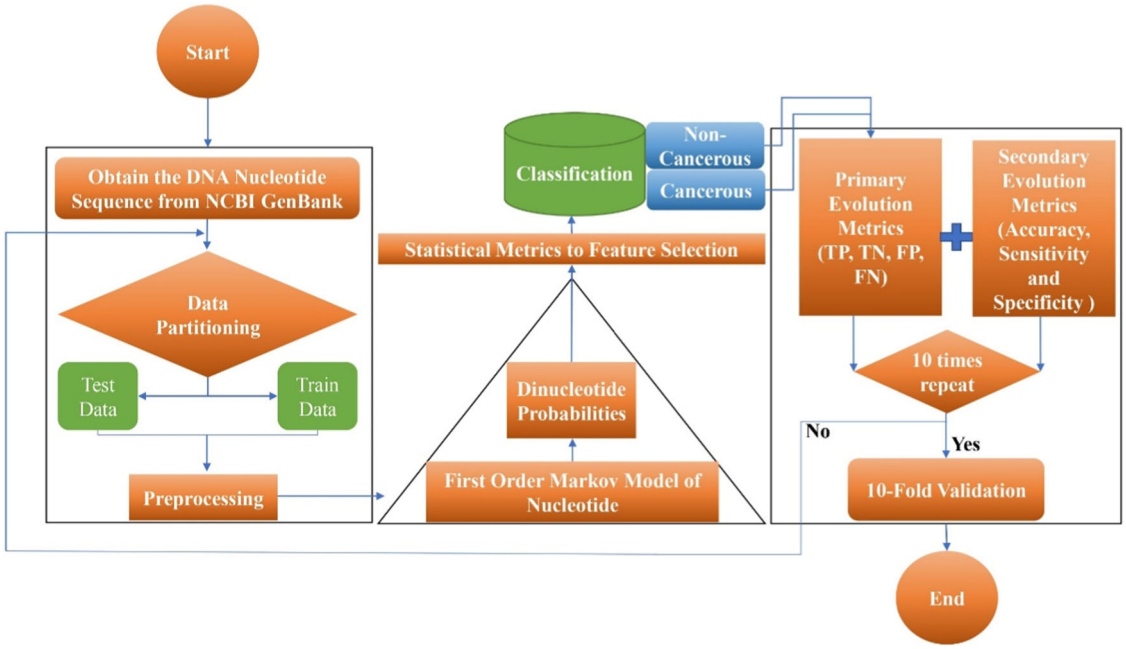

Fig. 3 Flowchart of Proposed Algorithms

Table 1 Recent Papers Case Study Details [4, 5, 15–20]

Sr. No	Disease	No. of Cancer	No. of Non-Cancer	Total
1	Breast [4, 5, 15–20]			
	(i) CDS	80	80	160
	(ii) N-CDS	77	80	157
	(iii) Whole Raw Sequence	80	80	160
Total	Breast Data	237	240	477
2	Colon [4, 5, 15–20]			
	(i) CDS	46	62	108
	(ii) N-CDS	46	55	101
	(iii) Whole Raw Sequence	46	62	108
Total	Colon Data	138	179	317
3	Prostate [4, 5, 15–20]			
	(i) CDS	52	55	107
	(ii) N-CDS	48	55	103
	(iii) Whole Raw Sequence	52	55	107
Total	Prostate Data	152	165	317
All Total	All Disease Total	527	584	1111

Applications of pattern recognition are found in many domains, such as intelligent system modeling and development [4, 31]. A pattern recognition model consists of four fundamental processes: feature extraction, feature selection, classification design, and evaluation [4, 22]. In the testing stage of these systems, the model's parameters are frequently established during training to classify test data [4, 22]. Feature extraction and feature selection in the classification step will be carried out using a first-order Markov model, which will be covered in the following section.

First-Order Markov Chain Model

X is a random variable that changes with the independent parameter *'t'*, sometimes known as a time parameter. For the stochastic variable X at time t, *'I'* stands for the collection of all possible states. If a stochastic process satisfies the criteria that follow, it is time-homogeneous [4, 23, 24].

$$P[X(t) \leq i | X(t_n = i_n)] = P[X(t - t_n) | X(0) = i_n] \quad (1)$$

Discrete-Time Markov Chain (DTMC): A stochastic process $\{X_t\}_{t\geq 0}$ is said to be a Markov chain (MC) if satisfied following condition [4, 23, 24]:

$$\begin{aligned} P(X_{n+1} = i_{n+1} | X_n = i_n, X_{n-1} = i_{n-1} \ldots X_0 = i_0) \\ = P(X_{n+1} = i_{n+1} | X_n = i_n) \end{aligned} \quad (2)$$

The conditional probability distribution of the system at a future stage only depends on its current state, not on its stage at a previous state. Assuming that the DTMC is time-homogenous, the transition probabilities result in a squared matrix known as a transition matrix when all states are considered. The characteristics of the time-

homogeneous DTMC's internal structure are represented by the transition matrix.

A time-homogeneous discrete-time Markov chain (DTMC) with a state-space I including the nucleotide symbols {A, C, G, T} and a discrete parameter space T is assumed to apply to a DNA string. To simulate the DNA sequence's characteristics in this case, a first-order Markov chain is utilized. This modeling technique considers each nucleotide at its place as a state. This methodology is used to analyze DNA sequences of any length [4, 31]. Every nucleotide is seen as a distinct state in its place to capture the first-order Markov chain properties. The random probability of finding a given nucleotide following the independent estimation of each kind of nucleotide. 16 values are calculated for each sample, which represents the differences in nucleotide count. To comprehend the features of the sequence, the modeling technique essentially considers DNA as a Markov chain, where the transition probabilities between nucleotides are computed [4, 31].

After a detailed study of the coding region, non-coding region, and the whole raw DNA sequence, the recommended approach successfully differentiates between cancerous and non-cancerous samples. DNA sequences are thoroughly analyzed resulting in a first-order Markov transition matrix for every nucleotide pair in the sample. Nucleotide pair occurrences in the sequences are computed and used to build the matrix. A complete transition probability matrix is obtained by computing a probability distribution for every pair of sixteen nucleotides in a DNA sequence. The Markovian nature of these transition probabilities is verified by comparing them to the results obtained by Amin Khodaei et al. [4]. A transformational shift that represents the conditional probability of sixteen nucleotide pairs occurring in a DNA sequence is used. Equation (3) calculates the matrix's elements, which need to be normalized. The resultant matrix is then normalized group-wise. Four distinct groups, each with a shared initial nucleotide, are formed from the sixteen matrix components. Lastly, dividing each group's values by their total is performed [4, 31]. For example, the matrix's cross-section between the fourth row and the second column. P(C | T) denotes the probability of event C in this case, given the possibility of event T. One by one, the remaining probability values in the matrix are calculated similarly.

$$Trans[M] = \begin{bmatrix} P_{A|A} & P_{C|A} & P_{G|A} & P_{T|A} \\ P_{A|C} & P_{C|C} & P_{G|C} & P_{T|C} \\ P_{A|G} & P_{C|G} & P_{G|G} & P_{T|G} \\ P_{A|T} & P_{C|T} & P_{G|T} & P_{T|T} \end{bmatrix} \quad (3)$$

The following Eq. (4) gives us another way to represent the previous Eq. (3),

$$Trans[M] = \begin{bmatrix} P_{AA} & P_{AC} & P_{AG} & P_{AT} \\ P_{CA} & P_{CC} & P_{CG} & P_{CT} \\ P_{GA} & P_{GC} & P_{GG} & P_{GT} \\ P_{TA} & P_{TC} & P_{TG} & P_{TT} \end{bmatrix} \quad (4)$$

The computational matrix format for dinucleotide patterns of chemical units seen in DNA sequences was found by using this process. Using the normalization approach, the transition matrix is converted into the distinct structure of the transition matrix of a Markovian chain. The transition matrix's overall probabilities merged into one by using this normalization process on each row. Ultimately, a classification is made of the Markovian transition matrix data. Regardless of the classification of a sample as non-cancerous or cancerous, the same process is used for it. Stated differently, the Markov model is utilized in the pattern recognition modeling process for feature extraction. Markov chains play a crucial role in the extraction and selection of characteristics. Statistical analysis is used to provide this feature, as will be discussed in more detail in the next section.

Kernel Logistic Regression (KLR)

A classification strategy is used in all pattern recognition models, and among these techniques, Logistic Regression (LR) is one of the most well-known techniques [4, 32]. Logistic regression is especially useful for binary dependent variables (binary classification) that have two classes: either elected or not, a policy adopted or not, a disease present or absent, and so on. It does this by combining a set of independent variables to effectively capture the variations in the dependent variable. Typically, an event is denoted by the code '1' for one category and '0' for the other [25, 26, 32].

In binary classification, the logistic regression (LR), where one group is labeled as $y = +1$, indicating represents cancerous DNA sample data, and the other as $y = 0$, showing non-cancerous DNA sequence. By correctly splitting data points into two categories, fitting a linear model to the input characteristics, and producing a probabilistic classification of the data points in datasets, LR attempts to estimate the probability of an event occurring. The following is the format of the LRM classification function [4, 25],

$$y = f(x) = \alpha + \beta x + \epsilon \quad (5)$$

Where y is the dependent variable attempting prediction (which is a of the class in sample data x), and x is the independent variable. This relationship is defined by the

equation $y = f(x)$. The value of y is shown by the intercept term (α) when x is equal to 0. At the same time, for every unit increase in x, the regression coefficient (β) measures the change in y and indicates related movements. Differences or residual variations in the model are represented by the stochastic term (ϵ) [25].

KLR is very effective in nonlinear classification because it estimates class-posterior probability using the log-linear function combination of the kernel. In the present model, a discriminant function used to solve classification problems is studied with specific focus on the role of the kernel function. The primary goal is to transform the original input space into a high-dimensional feature space. In this case, the kernel function plays a crucial role in carrying out a non-linear transformation on the input vector x, which is represented by dinucleotide patterns [35–38]. Thus, logistic regression (LR)'s nonlinear expression can be expressed as follows.:

$$f(x) = logit(p) = \vec{w}.\vec{x} + b \tag{6}$$

where w and b stand for the optimal model parameters that were obtained by minimizing a cost function, and $f(x)$ is used to determine the class of the sample data x. The regularized negative-log probability of the data is represented by this function. Furthermore, p denotes a probability associated with dinucleotide patterns. Following is the breakdown of how KRM classifies sample data set $D.S_{KRM}$ [4]:

$$D.S_{KLR} = \left[(x_i, y_i), x_i \in R^d, y_i \in \{0, 1\}\right] \tag{7}$$

Support Vector Machine (SVM)

Let's choose $y = +1$ in this case to represent data from cancerous DNA samples, and $y = -1$ to represent DNA sequences that are non- cancerous. According to this approach, Dataset D is considered to be linearly separable in a d-dimensional space if a hyperplane with coefficients w can efficiently divide the two sample data categories in the feature space. The SVM classification function denoted by $f(x)$ which is given in the following equation [4, 39],

$$f(x) = \vec{w}.\vec{x} + b \tag{8}$$

where $f(x)$ sign to identify the class of the sample data x. Following is the breakdown of how SVM classifies sample data set $D.S_{SVM}$ [4]:

$$D.S_{SVM} = \left[(x_i, y_i), x_i \in R^d, y_i \in \{-1, 1\}\right] \tag{9}$$

Table 2 Kernel Functions [4, 5, 32, 39, 42]

Kernel Name	Equation
Linear	$X_i^T.X_j$
Polynomial	$(X^T.X_i + 1)^P$
RBF	$e^{-\frac{1}{2\gamma^2}\lvert X - X_i\rvert^2}$
Sigmoid	$tanh\ (X^T.X_i + 1)$

To discriminate between classes, KRM and SVM requires a linear decision boundary in the feature space. However, if the data is not linearly separable in the original feature space, its performance can be changed. In these kinds of situations, many strategies are used to overcome this limitation. Using nonlinear functions to translate the data into a higher-dimensional space where a linear decision boundary is more useful is a popular technique. Depending on the particulars of the given situation, a variety of functions, including the polynomial functions, Radial Basis Function (RBF) kernel, and the Sigmoid kernel, can be used to perform this transformation. Table 2 is a comprehensive list of these transformation functions [4, 5, 39].

Evolution of Model Performance

Evaluation of model performance is critical in pattern recognition, particularly in classification tasks. The key metrics in the classification model for the present study are False Positive (FP), False Negative (FN), True Positive (TP), and True Negative (TN). These basic parameters may also be used to create a variety of secondary metrics like precision, sensitivity(recall), specificity, F1-score, and accuracy. In this work, we analyse and evaluate various classification algorithms based on the accuracy criterion. Precision, sensitivity(recall), specificity, F1-score, and accuracy are calculated in the following equation from 10 to 14 [4, 5, 7]:

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{11}$$

$$Specificity = \frac{TN}{TN + FP} \tag{12}$$

$$\text{F1} - \text{Score} = 2\left(\frac{Recall \times Precision}{Recall + Precision}\right) \tag{13}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{14}$$

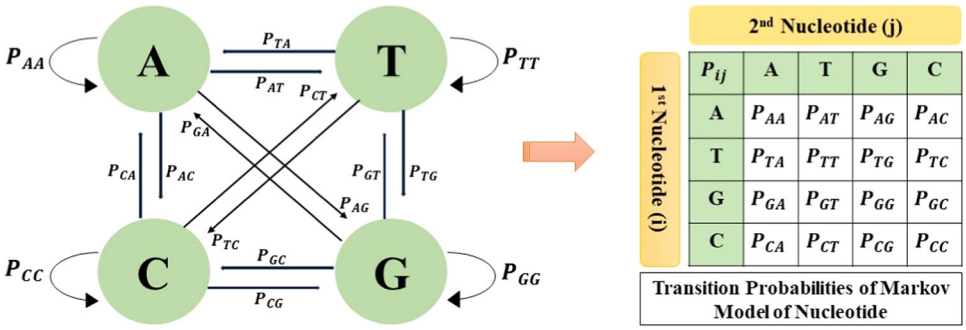

Fig. 4 1st-Order Markov Model of Nucleotides with Transition Probabilities

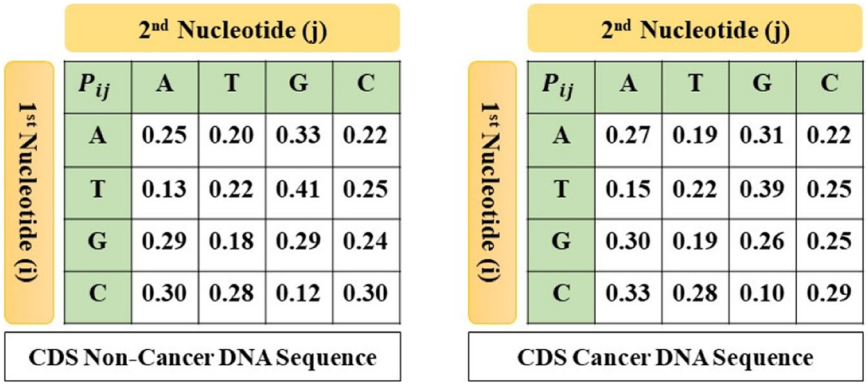

Fig. 5 Transition Probabilities of Breast CDS Region for Non-Cancer and Cancer

Cross Validation of Classification Model

The validation test is a widely used and crucial method for evaluating model performance in the fields of pattern recognition and classification. K-Fold Cross-Validation is one method that is frequently applied in this field. Using this approach, the dataset is divided into K folds or subsets, and the model is evaluated on the remaining fold after repeatedly training on K-1 folds. The comprehensive evaluation is obtained by averaging the performance measures for a total of K iterations [4]. The challenge of generalization capacity is addressed by this technique, specifically about the size of training data. The model's capacity to generalize might be limited by a decrease in the amount of training data. At the same time, the size of the test data compared to the total dataset tends to improve error estimates in classification. K-Fold Cross-Validation overcomes such problems and provides a more accurate evaluation of the model's performance in different scenarios by systematically testing the model over several folds [4, 27, 28].

Result and Discussion

To analyze and simulate the results Python is used. Figure 4 represents the 16 pairing of nucleotide (dinucleotide) transition probabilities for the four states A, G, T, and C of the first-order Markov chain model of nucleotides. Additionally, Fig. 4 represents the samples from the prostate, colon, and breast that are linked to certain genes. According to the proposed first-order Markov model of nucleotides, which applies to all disease samples. The probability matrix of the first-order Markov chain of nucleotides for CDS of Breast disease mentioned is shown in Fig. 5. The transition matrix entries in this figure are rounded to two decimal places. As an illustration, the element in row 2 and column 3 has an average probability of 0.41 in non-cancer and the element in row 2 and column 3 has an average probability of 0.39 in cancer which indicates that G nucleotides will typically come after T nucleotides. In the same way, for the remaining two N-CDS regions and the whole raw sequence, transition probabilities of dinucleotides are constructed from Fig. 5. To analyze nucleotide distributions in prostate and colon cancer covering CDS, N-CDS, Whole raw sequence, we currently use a first-order Markov model. As a result, transition probabilities are computed; however, they are not explicitly described here. Figure 5 makes it clear that each row's overall probability is equal to 1. Using the selected database, the described technique first builds a transition matrix for each instance in the case study, including samples that are cancerous and non-cancerous. The feature, which consists of 16 features that represent dinucleotides, is seen as a row-by-row representation of the transition matrix.

The resulting features have a wide range of applications, such as separation and classification. An in-depth

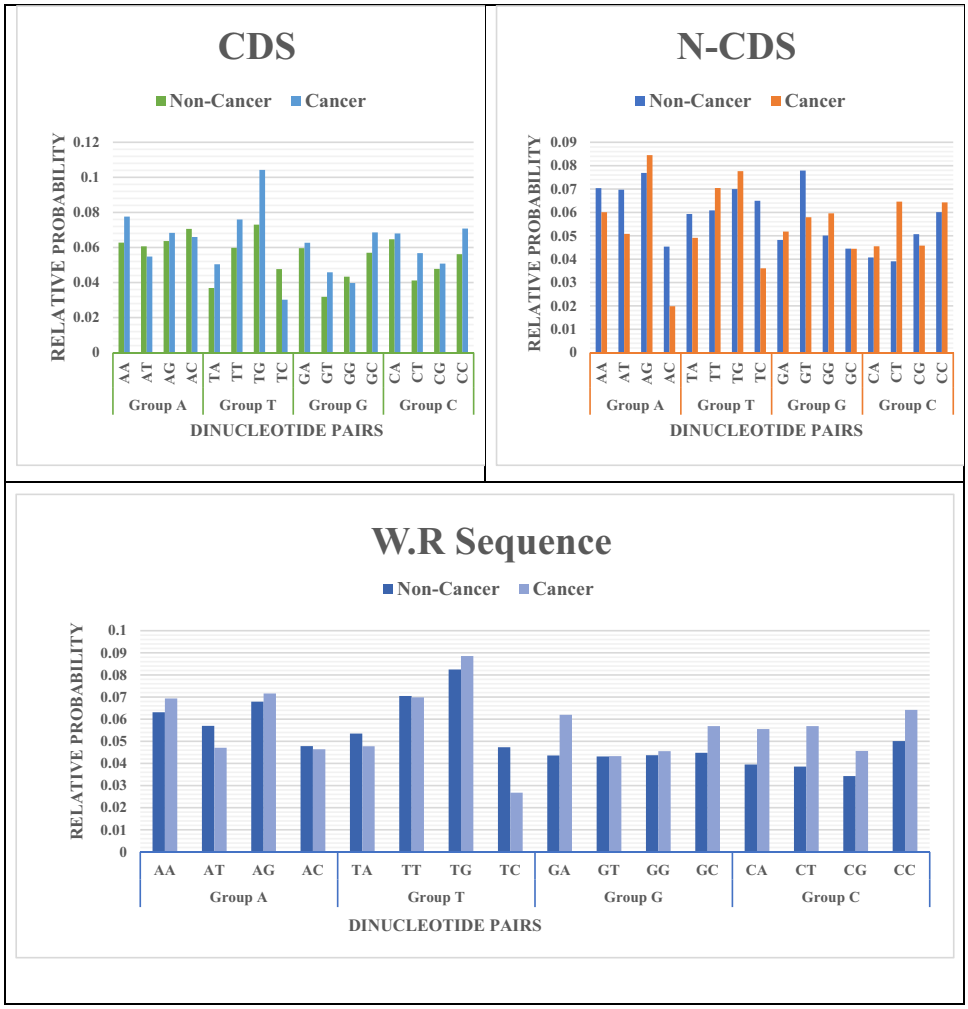

Fig. 6 MAD of Conditional Probability of Breast Disease

understanding of differentiating features can be obtained by looking at each element in the matrix. The transition matrix elements with MAD (Mean Absolute Deviation) are displayed in Figs. 6–8 for samples with and without cancer across three different diseases. In Figs. 6–8, four different groups can be observed on the horizontal axis, each representing a single nucleotide that acts as the first element in dinucleotide pairs. The vertical axis shows the relative frequency of recorded data by utilizing the group normalization approach. All three diseases show significant differences in the statistical metrics for DNA sequences associated with cancerous and non-cancerous indications, as shown in Figs. 6–8.

In the case of breast disease, non-cancerous samples had a higher chance of recognizing a particular dinucleotide pattern TC, AT, AC, and GG in the CDS region than cancerous samples in Fig. 6. Additionally, the probability of AT, AA, AC, TC, and GT in the N-CDS region is higher in non-cancerous samples than in cancerous ones. Furthermore, all samples in the raw sequence show that non-cancerous samples had greater probabilities of TC, TA, and AT than cancerous ones. The probability of detecting a particular dinucleotide pattern in non-cancerous samples is higher than in cancerous ones, as shown in Figs. 7 and 8 for colon and prostate diseases, respectively. To put it another way, a threshold value can be defined to differentiate between data that is cancerous and non-cancerous. This idea is a framework for the idea of a pre-processing technique that uses statistical metrics that have significant variations to differentiate between DNA sequences that are cancerous and those that are not.

The study of the MAD to identify significant differences between cancerous and non-cancerous groups demonstrates that our data is classifiable. Using a discriminative phase, whereby an appropriate machine-learning algorithm makes use of the statistical features for sample classification, can efficiently accomplish this classification. Both SVM classifiers and KLR provide appropriate methods for handling fundamental relationships and non-linear relationships among features. The application of statistical features in KLR creates an optimum decision boundary, highlighting the significance of model selection for accurate classification. Similarly, SVM classifiers use feature space's

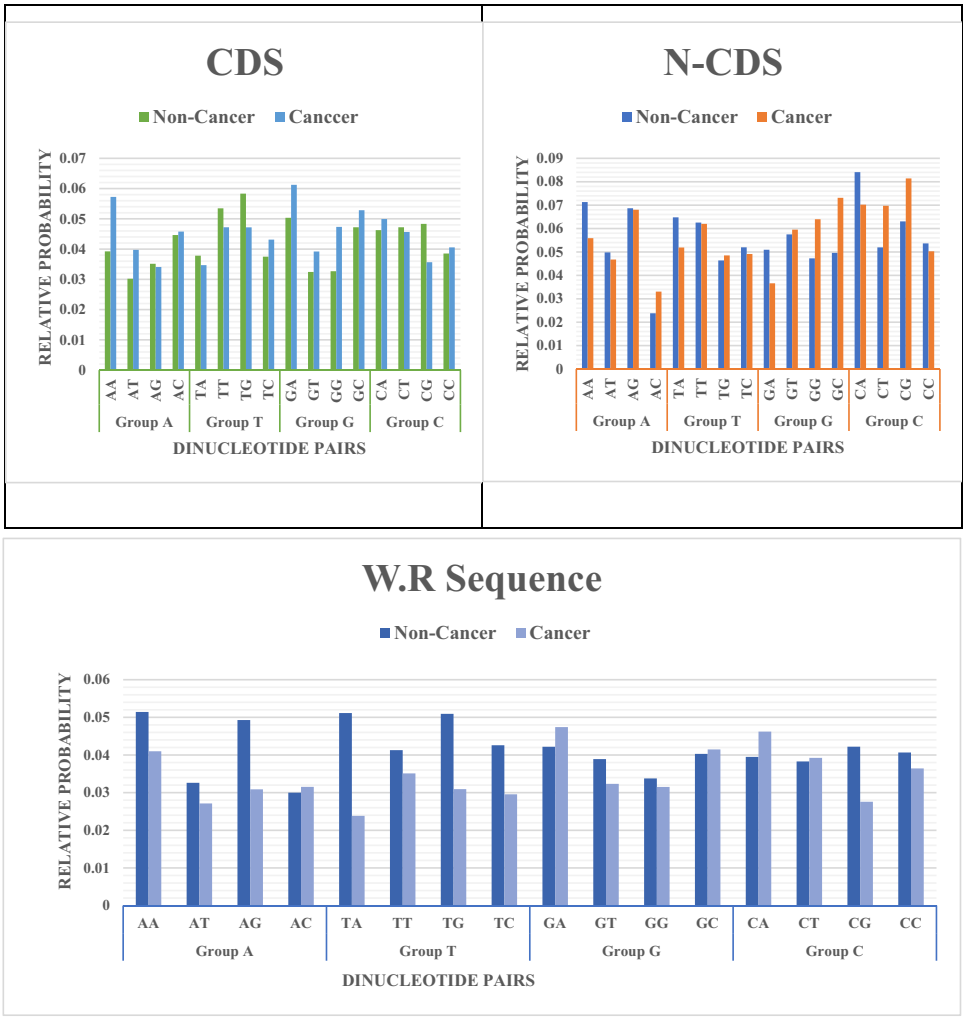

Fig. 7 MAD of Conditional Probability of Colon Disease

statistical features to create the best classification hyperplanes; selecting an appropriate kernel function for data classification based on available features is a crucial factor.

In the present study, SVM and KLR classifiers have been applied with various kernel functions on a feature space of 338 DNA sequence data. Coding DNA Sequences (CDS) with 48 cases in breast data, 33 instances in colon data, and 33 cases in prostate data make up these sequences. Furthermore, 48 cases of Non-Coding DNA Sequences (N-CDS) from breast data, 31 cases from colon data, and 31 cases from prostate data are included. Additionally, whole raw DNA sequences (W.R Sequences) are included in the study; there are 48 instances in the data related to the breast, 33 cases in the colon, and 33 cases in the prostate. For both SVM and KLR classifiers, the performance of the classifiers was evaluated by comparing their various kernel functions.

Figures 9–11 present a thorough comparison using the TP, FN, FP, and TN criteria for both SVM and KLR classifiers. This investigation tested several kernel functions from Tables 3–5 to provide insight into how well they performed. It also illustrates the classification accuracy attained by various kernel functions. The testing and comparison studies that follow validate the successful use of these kernel functions in the classification method.

In Figs. 9–11, the learning approach names for classification kernels are shown on the horizontal axis. The number of classification values for 177 cancerous samples and 161 noncancerous samples is shown on the vertical axis of Figs. 9–11 using the metrics TP, FN, FP, and TN. The noncancerous samples include CDS, N-CDS, and the whole raw sequence for the three diseases (breast, colon, and prostate), whereas the cancerous samples include CDS, N-CDS, and the whole raw sequence. The distribution of cancerous samples for breast disease is as follows: CDS: 25, N-CDS: 23, and the whole raw sequence: 25. On the other hand, CDS: 23, N-CDS: 25, and the whole raw sequence: 23 make up noncancerous samples for breast disease. Figures 10 and 11, which show the distribution of cancerous and noncancerous samples, show similar trends for the other two diseases, prostate and colon. The effective performance of the SVM and KLR kernel functions in handling nonlinear feature spaces can be observed in this figure. To

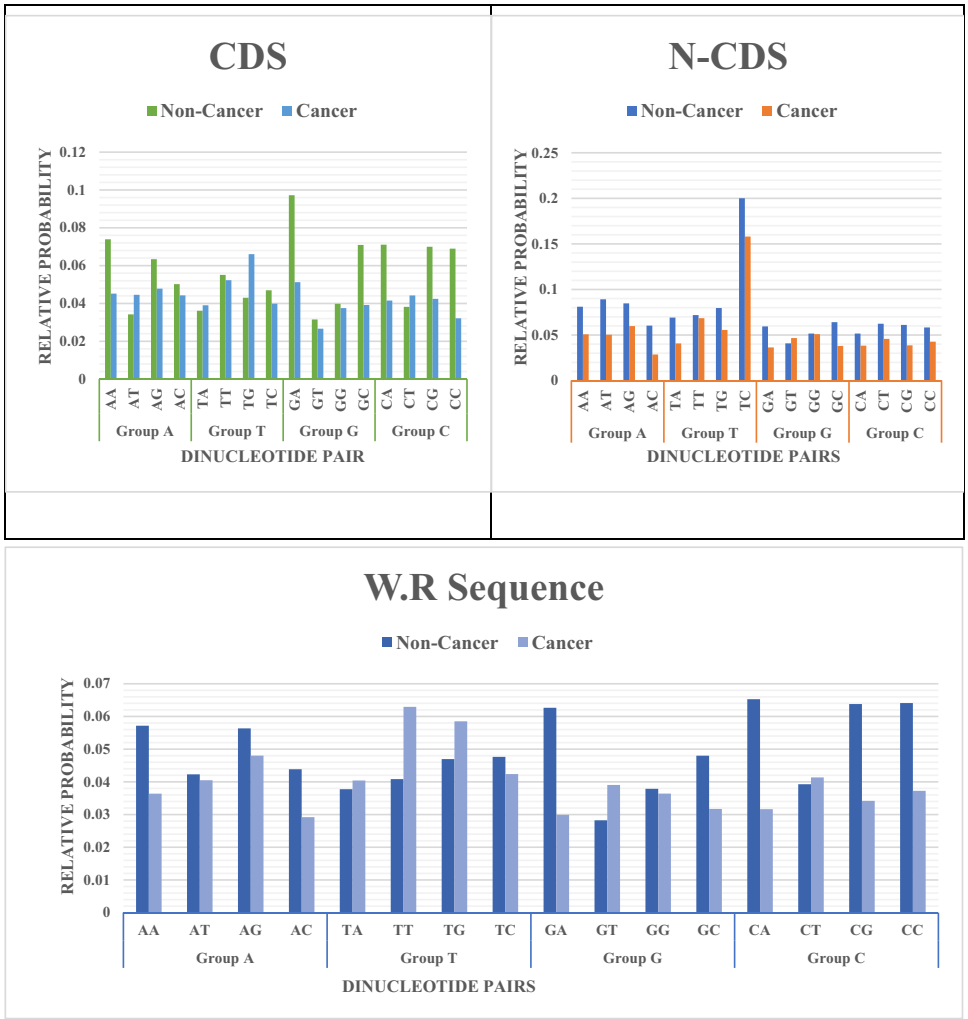

Fig. 8 MAD of Conditional Probability of Prostate Disease

highlight the best outcomes and highlight their better performance, kernel functions like polynomial and RBF in KLR and polynomial, RBF, and Sigmoid in SVM are used in both analyses.

Additionally, Tables 3–5 demonstrate the performance of SVM classification and KLR techniques using different kernel functions. Based on specified criteria for performance, evaluation is carried out in three distinct regions related to prostate, colon, and breast diseases. A 10-fold cross-validation technique and an identified accuracy-based criterion are used to assist the comparison. The criteria that are provided include performance metrics and accuracy, and they outline the results for each of the three regions. Classification outcomes are more accurate when accuracy scores are higher.

Tables 3–5 show that for SVM and KLR for all three diseases, a linear function produces better classification accuracy in some regions (CDS, N-CDS, and W.R Sequence). Additionally, for all three diseases, polynomial, RBF, and sigmoid kernel functions show improved classification accuracy for both SVM and KLR in some regions (CDS, N-CDS, and W.R Sequence). It is insufficient to evaluate the success of a method based just on a single evaluation of accuracy-based criteria. In this paper, we use a widely used method with regular evaluations to overcome the typical limitations of machine learning challenges. To verify the effectiveness of our suggested approach, we use the K-Fold methodology in our experiment. Tables 3–5 present the results of dimension reduction and classification after 10 rounds ($K = 10$).

Tables 3–5 show the outcomes of our comparison between the Support Vector Machine (SVM) and Kernel Logistic Regression (KLR) with kernels function. In breast disease, Table 3 shows that the RBF, Polynomial2, and Linear kernels were quite accurate in SVM, while the KLR Polynomial and RBF kernels were very accurate in the CDS region. Table 3 shows that while Linear, Polynomial 2, and RBF demonstrated significant accuracy in the SVM, Polynomials 2 and 3 attained outstanding accuracy in the KLR for the N-CDS region. Polynomials 2, 3, and RBF

Fig. 9 Analysis of Breast Disease: A Comparison of Classification Kernels for KLR and SVM

showed significant accuracy in the KLR in the W.R raw sequence, whereas Linear, Polynomial 2, and RBF showed excellent accuracy in the SVM model (Table 3). In the context of colon disease (Table 4), polynomials 2 and 3 in the SVM model show significant accuracy. Likewise, polynomials 2 and 3 in the KLR show high accuracy in the CDS region. Additionally, polynomials 2 and 3 show outstanding accuracy in the KLR for the N-CDS region, according to Table 4, whereas linear, sigmoid, and RBF demonstrate notable accuracy in the SVM. In summary, Table 4 shows that, in the W.R raw sequence, polynomials 2 and 3 provide significant accuracy in the KLR, whereas linear, polynomial 2, and RBF indicate excellent accuracy in the SVM model. The linear, polynomial 2, and RBF models for prostate disease show impressive SVM performance in the CDS region (Table 5). Furthermore,

Fig. 10 Analysis of Colon Disease: A Comparison of Classification Kernels for KLR and SVM

polynomials 2 and 3 in the KLR show high accuracy in the CDS region. Table 5 reveals that in the N-CDS area, polynomials 2 and 3 show remarkable accuracy in the KLR, whereas linear, polynomials 2, 3, and RBF demonstrate significant accuracy in the SVM. Additionally, polynomials 2 and 3 demonstrate good accuracy in the SVM model and significant accuracy in the KLR in the W.R raw sequence.

From the above discussion, when the linear kernel function was used for all three diseases, the experimental results showed that it performed less accurately in some regions (CDS, N-CDS, and W.R Sequence) than other kernel functions. The lack of linear separability among the sample points in the new feature space is shown by the accuracy issues seen with linear kernel functions. As such, using non-linear classifiers is crucial to getting better accuracy.

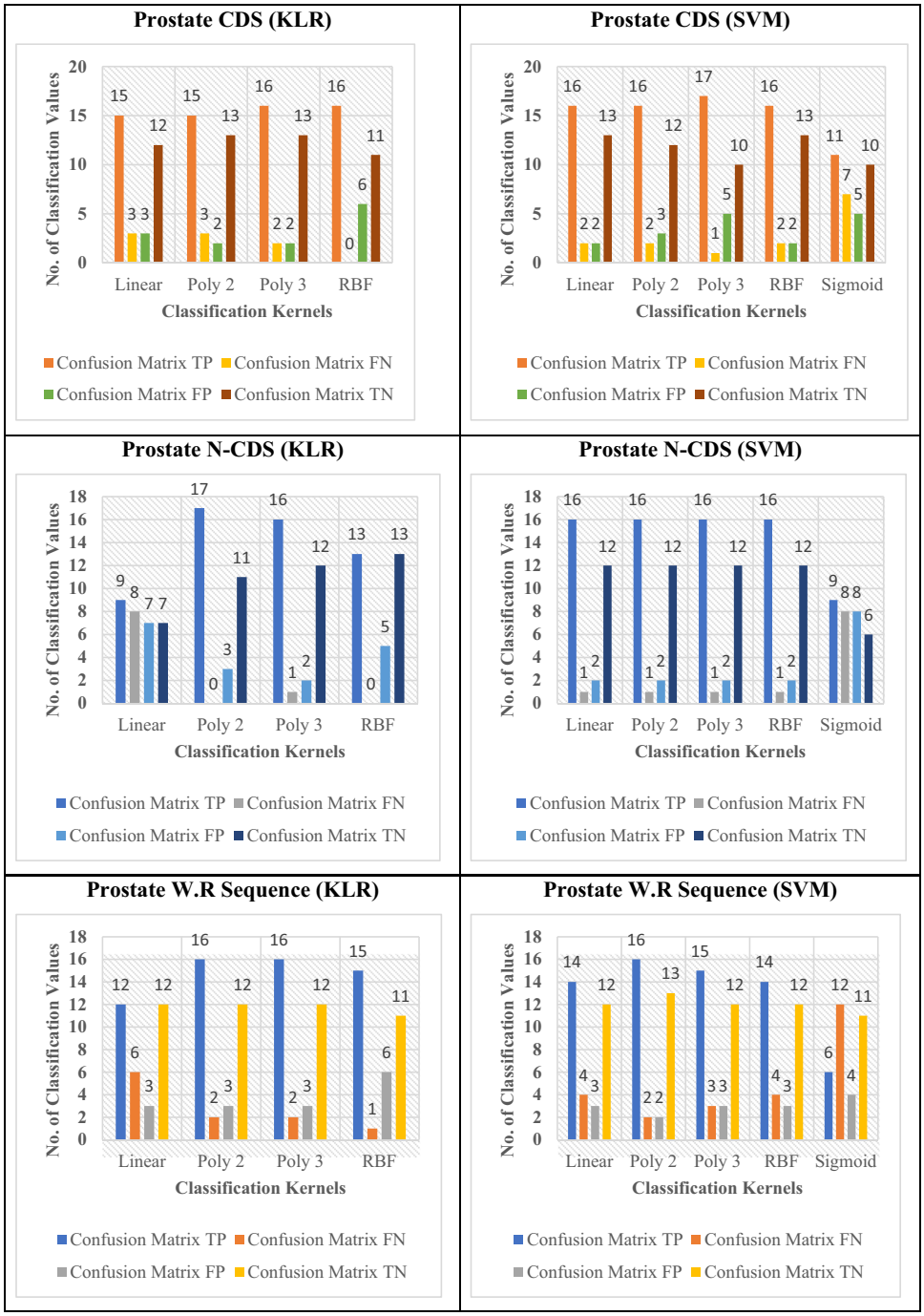

Fig. 11 Analysis of Prostate Disease: A Comparison of Classification Kernels for KLR and SVM

Discussion

In addition to the evaluation of the chosen dataset, the suggested methodology is used for a dataset utilized in previous studies [4, 5, 15–20]. The performance of the proposed approach when used with these datasets is also covered in detail in this section. For example, Fig. 12 shows the transition probabilities obtained from the first-order Markov chain of nucleotide as features. More specifically, Fig. 12 shows the average probability of dinucleotide pairs observed in cancerous and non-cancerous samples in each of the three regions related to breast disease. In the context of colon and prostate diseases, respectively, Figs. 13 and 14 show the average probability of dinucleotide pairs observed in samples that are cancerous and non-cancerous within each of the three regions. All 16 dinucleotide pairs appear on the vertical axes of Figs. 12–14, while the horizontal axes show the grouped normalized values of the percentages of each nucleotide's occurrence frequencies. The transition probabilities of non-cancerous cases over 16 pairs

Table 3 Analyzing classification methods for Breast Disease with a focus on accuracy and the 10-Fold criterion in evaluating KLR and SVM

Breast CDS (KLR)			Breast CDS (SVM)		
Kernels	Accuracy	Performance	Kernels	Accuracy	Performance
Linear	0.5625	0.63333333	Linear	0.8958333	0.8765152
Polynomial 2	1	0.9734848	Polynomial 2	0.875	0.8931818
Polynomial 3	0.875	0.9560606	Polynomial 3	0.6875	0.8659091
RBF	0.9583333	0.9189394	RBF	0.8958333	0.8939394
			Sigmoid	0.4791667	0.5
Breast N-CDS (KLR)			**Breast N-CDS (SVM)**		
Kernels	**Accuracy**	**Performance**	**Kernels**	**Accuracy**	**Performance**
Linear	0.625	0.6063636	Linear	0.9375	0.8081818
Polynomial 2	0.875	0.88	Polynomial 2	0.8125	0.7818182
Polynomial 3	0.9791667	0.89	Polynomial 3	0.7708333	0.7172727
RBF	0.7291667	0.7790909	RBF	0.9375	0.8081818
			Sigmoid	0.4583333	0.5327273
Breast W.R Sequence (KLR)			**Breast W.R Sequence (SVM)**		
Kernels	**Accuracy**	**Performance**	**Kernels**	**Accuracy**	**Performance**
Linear	0.6875	0.6962121	Linear	0.85417	0.8393939
Polynomial 2	0.9583333	0.9204545	Polynomial 2	0.875	0.8590909
Polynomial 3	0.8958333	0.9007576	Polynomial 3	0.79167	0.7325758
RBF	0.8958333	0.8128788	RBF	0.85417	0.8484848
			Sigmoid	0.39583	0.4727273

Table 4 Analyzing classification methods for Colon Disease with a focus on accuracy and the 10-Fold criterion in evaluating KLR and SVM

Colon CDS (KLR)			Colon CDS (SVM)		
Kernels	Accuracy	Performance	Kernels	Accuracy	Performance
Linear	0.8484848	0.8910714	Linear	0.7575758	0.9214286
Polynomial 2	0.8787879	0.9196429	Polynomial 2	0.8484848	0.8535714
Polynomial 3	0.9090909	0.9589286	Polynomial 3	0.8181818	0.8285714
RBF	0.8484848	0.8767857	RBF	0.7575758	0.9214286
			Sigmoid	0.6363636	0.7982143
Colon N-CDS (KLR)			**Colon N-CDS (SVM)**		
Kernels	**Accuracy**	**Performance**	**Kernels**	**Accuracy**	**Performance**
Linear	0.7741935	0.6857143	Linear	0.7419355	0.6857143
Polynomial 2	0.8387097	0.8428571	Polynomial 2	0.6774194	0.6428571
Polynomial 3	0.7741935	0.8428571	Polynomial 3	0.6451613	0.5714286
RBF	0.5483871	0.6714286	RBF	0.7419355	0.6857143
			Sigmoid	0.8064516	0.6571429
Colon W.R Sequence (KLR)			**Colon W.R Sequence (SVM)**		
Kernels	**Accuracy**	**Performance**	**Kernels**	**Accuracy**	**Performance**
Linear	0.6363636	0.875	Linear	0.72727	0.8125
Polynomial 2	0.8787879	0.9071429	Polynomial 2	0.69697	0.8785714
Polynomial 3	0.8787879	0.9464286	Polynomial 3	0.75758	0.7196429
RBF	0.8181818	0.8107143	RBF	0.72727	0.8125
			Sigmoid	0.63636	0.7946429

of nucleotides show a significant connection to those in cancerous cases in all three regions, as seen in Figs. 12–14. Furthermore, one may interpret the distribution of average transition probabilities in different ways. The observed changes in both increased and reduced variations are related to the base genetic mutations seen in cancerous cells. Based on the discussion above, a practical threshold number or values can be determined for each classification. It is also

Table 5 Analyzing classification methods for Prostate Disease with a focus on accuracy and the 10-fold criterion in evaluating KLR and SVM

Prostate CDS (KLR)			Prostate CDS (SVM)		
Kernels	Accuracy	Performance	Kernels	Accuracy	Performance
Linear	0.8181818	0.7928571	Linear	0.8787879	0.9178571
Polynomial 2	0.8484848	0.9053571	Polynomial 2	0.8484848	0.8625
Polynomial 3	0.8787879	0.9482143	Polynomial 3	0.8181818	0.8214286
RBF	0.8181818	0.7696429	RBF	0.8787879	0.9178571
			Sigmoid	0.6363636	0.7571429
Prostate N-CDS (KLR)			**Prostate N-CDS (SVM)**		
Kernels	**Accuracy**	**Performance**	**Kernels**	**Accuracy**	**Performance**
Linear	0.516129	0.7375	Linear	0.9032258	0.7660714
Polynomial 2	0.9032258	0.8660714	Polynomial 2	0.9032258	0.7910714
Polynomial 3	0.9032258	0.8928571	Polynomial 3	0.9032258	0.8053571
RBF	0.8387097	0.875	RBF	0.9032258	0.7660714
			Sigmoid	0.483871	0.6928571
Prostate W.R Sequence (KLR)			**Prostate W.R Sequence (SVM)**		
Kernels	**Accuracy**	**Performance**	**Kernels**	**Accuracy**	**Performance**
Linear	0.7272727	0.7571429	Linear	0.78788	0.825
Polynomial 2	0.8484848	0.9196429	Polynomial 2	0.87879	0.8660714
Polynomial 3	0.8484848	0.8910714	Polynomial 3	0.81818	0.825
RBF	0.7878788	0.8678571	RBF	0.78788	0.825
			Sigmoid	0.51515	0.6107143

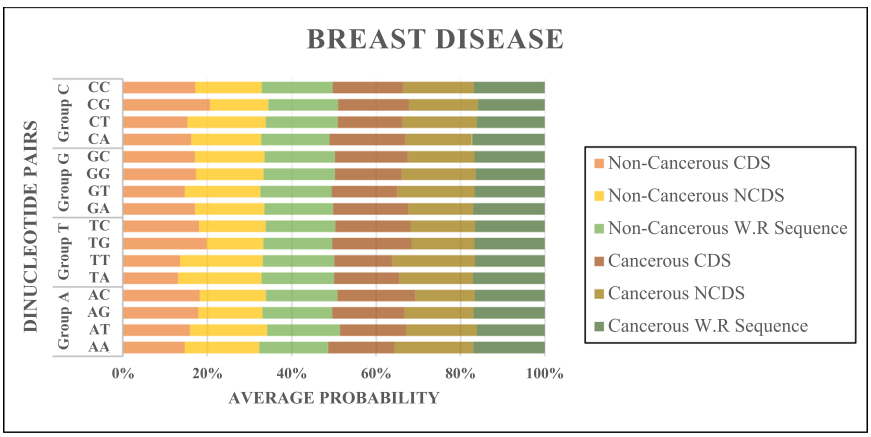

Fig. 12 Probability Distribution of Dinucleotide for CDS, N-CDS and W.R Sequence of Breast Disease

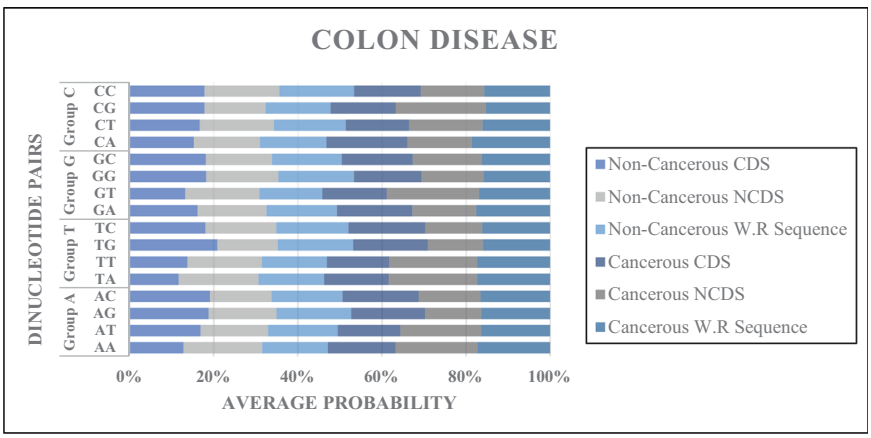

Fig. 13 Probability Distribution of Dinucleotide for CDS, N-CDS and W.R Sequence of Colon Disease

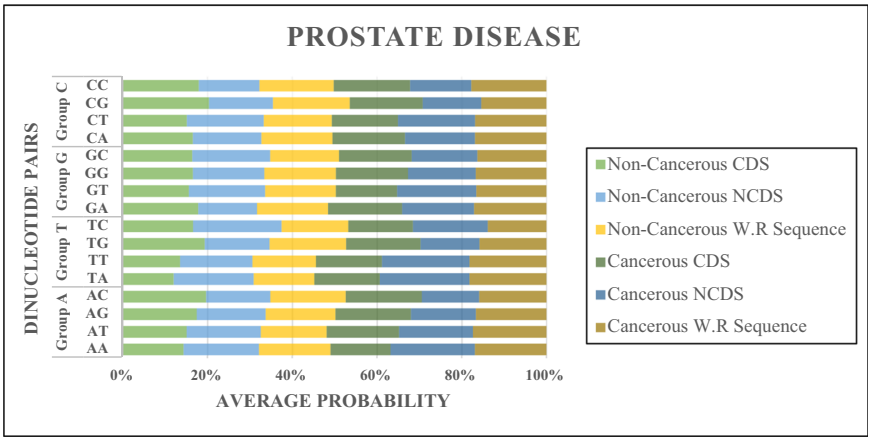

Fig. 14 Probability Distribution of Dinucleotide for CDS, N-CDS and W.R Sequence of Prostate Disease

observed that the outcomes are dependent on the particular cancer type and the genes associated with it. Taking a detailed look at each individual component might lead to a number of studies and discussions.

In particular, when considering situations involving cancer and non-cancer, precision, recall, and F1-score are essential secondary metrics for evaluating the effectiveness of classification algorithms. Among all presented cases under each classification, precision measures how well the algorithm identifies true instances whether or not they are cancerous. Conversely, recall measures the accuracy of the model in detecting actual cases inside each meaningful classification whether or not they are cancerous. In situations of cancer as well as non-cancer, finding a balance between recall and precision is important. One unique and useful measurement that offers a comprehensive evaluation of model performance is the F1-score, which looks at both recall and precision at the same time. According to our study, Figs. 15–17 shows the precision, recall, and F1-score for cases with and without cancer about three different diseases: the prostate, colon, and breast.

The vertical axis in Figs. 15–17 depicts the classification score, while the horizontal axis shows the secondary classification metrics which are precision, recall, and F1-Score for both cancer and non-cancer cases. In the context of the previous discussion, we compared the performance of Kernel Logistic Regression (KLR) and Support Vector Machine (SVM) using several kernels, including Linear, Polynomial (degrees 2 and 3), RBF, and Sigmoid. A brief of the Precision, Recall, and F1-score for the 'Cancer' and 'Non-Cancer' classes is provided in this comparison analysis.

Figure 15 represents a breast disease with CDS, N-CDS, and W.R. Sequence. It shows that the SVM's linear kernel outperforms the KLR for the cancer class in the CDS region in terms of both precision and recall. Furthermore, obtaining the highest possible precision, recall, and F1-score in KLR and SVM with Polynomial kernels both show perfect performance. When compared to SVM and KLR with an RBF kernel notably shows greater recall for the cancer class. On the other hand, the SVM using Sigmoid kernels for the cancer class has poor F1-score, recall, and precision. In the N-CDS area, Fig. 15 indicates that an SVM with a Linear kernel achieves higher precision, recall, and F1-score for the cancer class when compared to KLR. In terms of precision and recall, KLR is better than SVM, especially when it comes to Polynomial kernels. With the RBF kernel, SVM and KLR both function similarly perform. The cancer class identification precision of SVM appears to be limited, as seen by its low performance when using the Sigmoid kernel. The comparison of precision, recall, and F1-score for the cancer class in Fig. 15, as shown in the W.R Sequence, shows that SVM with a linear kernel performs better than KLR. Both SVM and KLR perform extremely well with Polynomial kernels in terms of F1-score, recall, and precision. The analysis of RBF kernels shows that KLR and SVM perform similarly and significantly, with good precision, recall, and F1-score. But, particularly in SVM, the Sigmoid kernel shows very little efficacy.

A colon disease implementing the CDS, N-CDS, and W.R. Sequence is depicted in Fig. 16. Better precision, recall, and F1-score for the cancer class are obtained using KLR with a Linear kernel for the CDS region when compared to SVM. With polynomial kernels, both SVM and KLR result in balanced results for cancer classification; still logistic regression maintains a slightly higher performance level. Similar results are obtained using RBF kernels for KLR and SVM. On the other hand, the recall, F1 score, and precision are all greater with the sigmoid kernel. In the N-CDS region, SVM performs well in recall whereas KLR with a Linear kernel shows greater precision. For both SVM and KLR, polynomial kernels show a trade-off between recall and precision. Remarkably, the RBF kernel results show that SVM outperforms KLR in terms of precision, recall, and F1-score for the cancer class, indicating a significant performance difference between the two models.

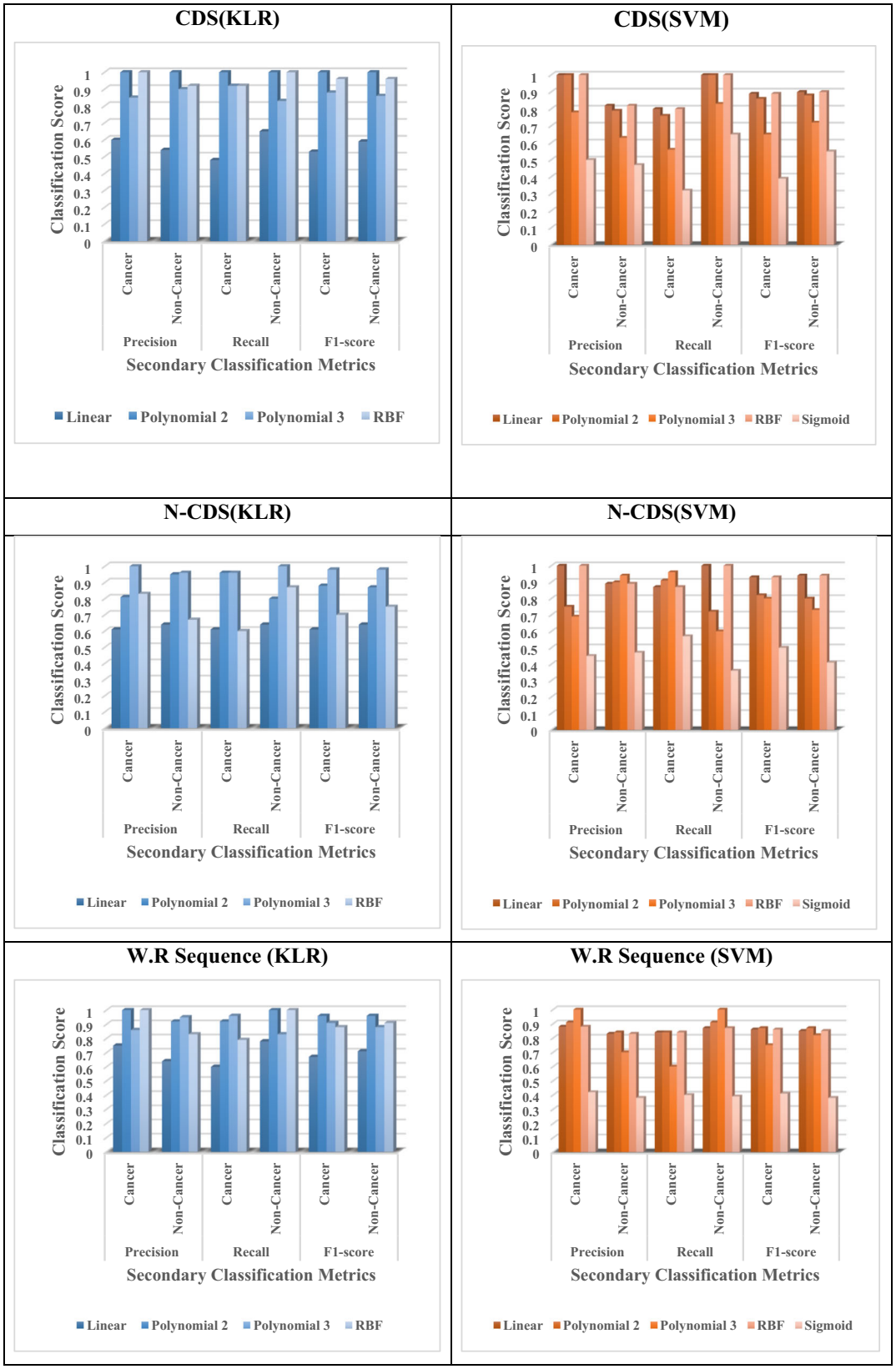

Fig. 15 Precision, Recall, and F1-Score as Secondary Classification Measures for Breast Disease along with the Classification Score

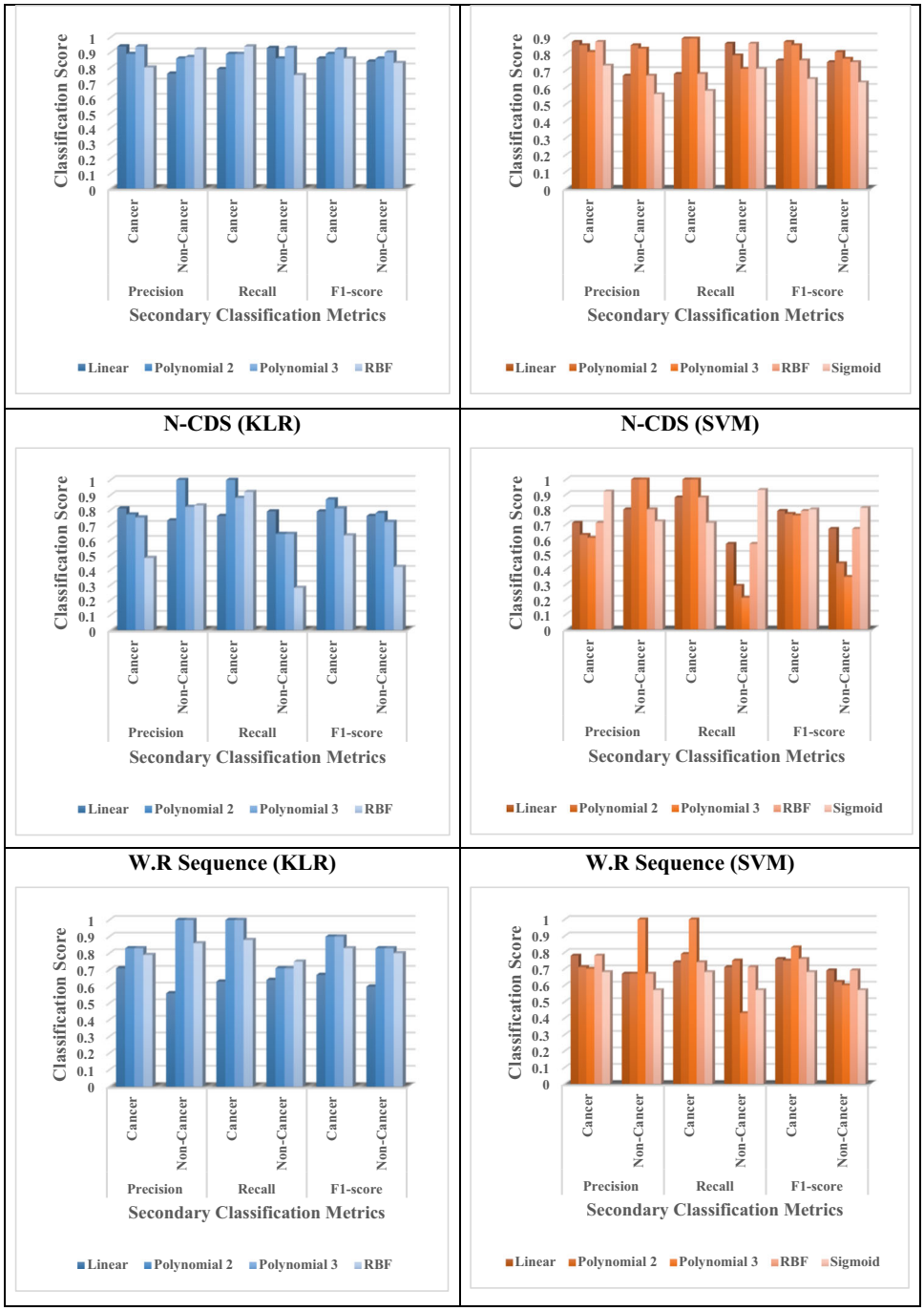

Fig. 16 Precision, Recall, and F1-Score as Secondary Classification Measures for Colon Disease along with the Classification Score

Furthermore, the Sigmoid kernel consistently achieves great performance in terms of F1 score, recall, and precision. For the cancer class in the W.R. Sequence, SVM with a linear kernel outperforms KLR in terms of F1 score, precision, and recall. With Polynomial kernels, KLR consistently provides good results, but SVM has inconsistent performance. When using RBF kernels, KLR and SVM perform similarly, attaining balanced precision, recall, and F1-score, whereas SVM performs well in terms of precision, recall, and F1-score.

In the same manner, a comparison of cancer and non-cancer classifications follows precision, recall, and F1-score for the kernel functions. Figure 17 depicts this analysis, which applies SVM and KLR models to prostate disease.

Furthermore, accurately identifying cancer cases detected by the proposed method depends significantly on this comparison analysis. Consequently, we have conducted a comparison analysis utilizing KLR and SVM to evaluate the efficacy of the suggested first-order Markov Model of nucleotides in classifying actual cancer cases. Using the

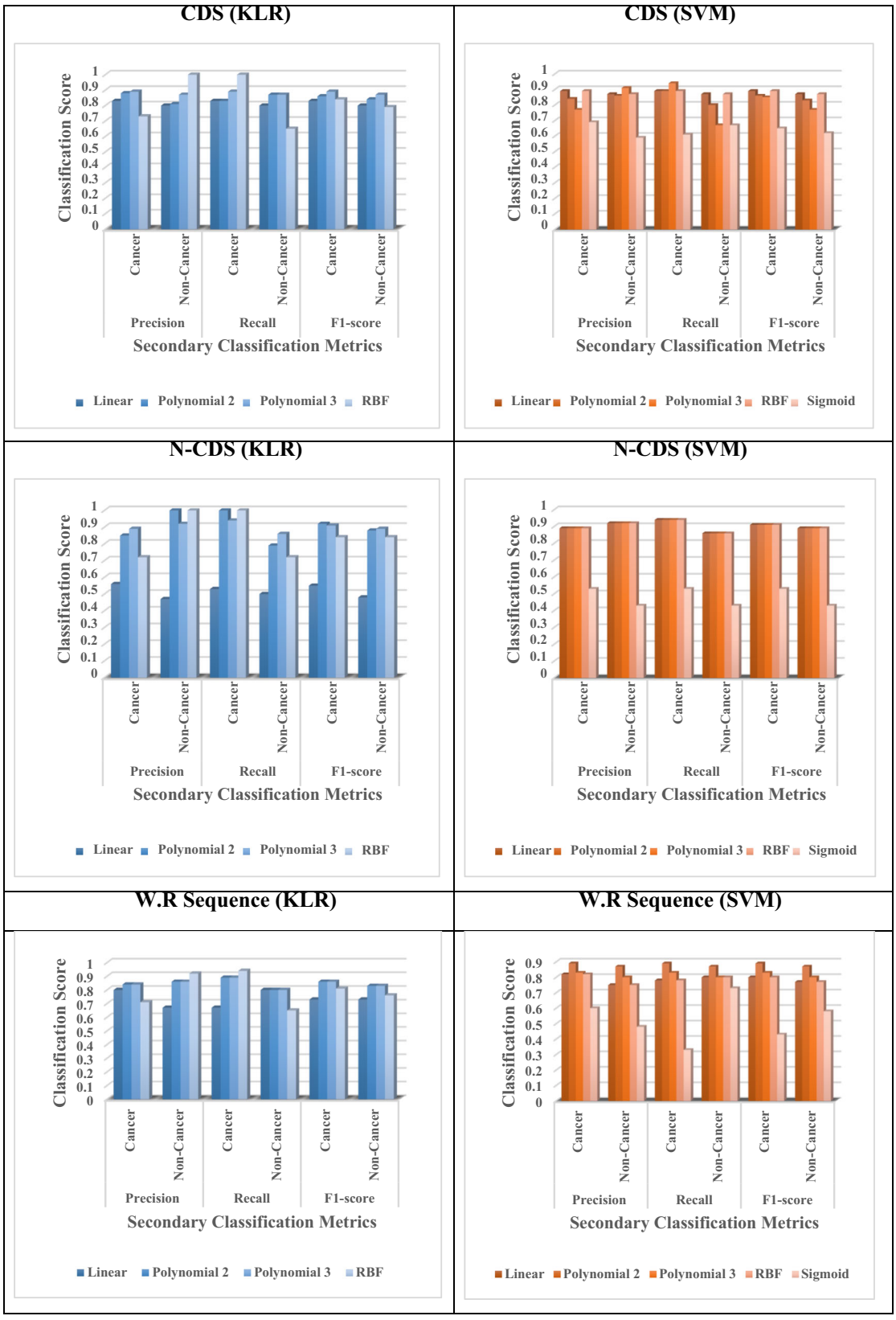

Fig. 17 Precision, Recall, and F1-Score as Secondary Classification Measures for Prostate Disease along with the Classification Score

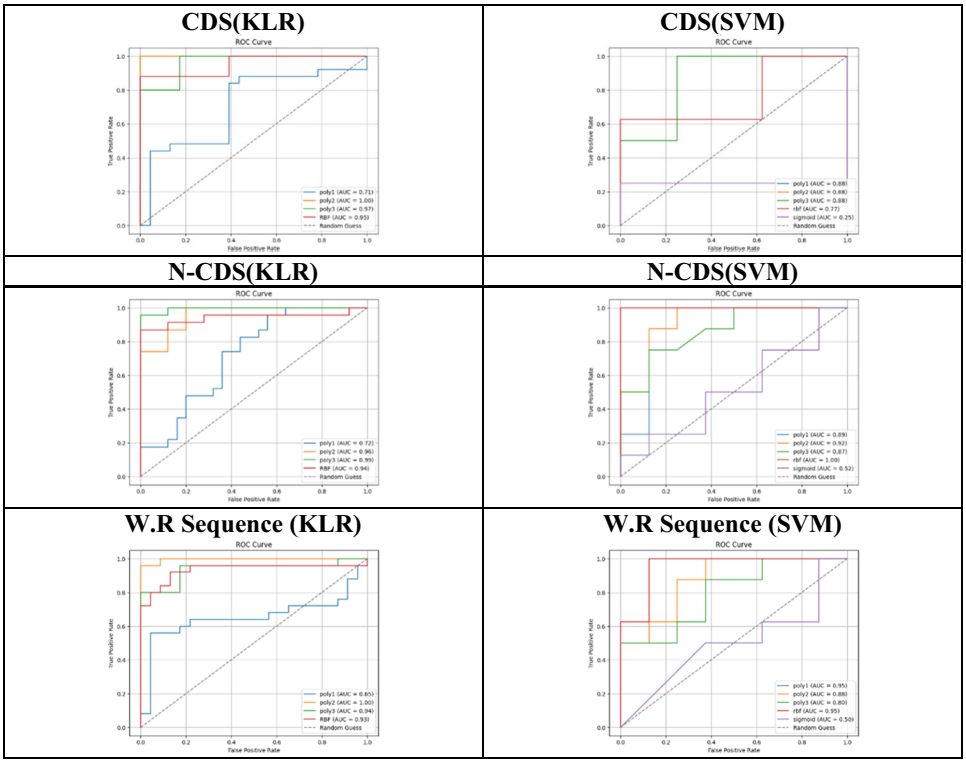

Fig. 18 Comparative ROC Curve Analysis of CDS, N-CDS, and W.R Sequence Regions in Breast Disease

Receiver Operating Characteristic (ROC) curve, the data was visualised to evaluate test performance and improve sensitivity and specificity. It was emphasized how important ROC curves are, especially in medical settings like cancer detection. Figures 18–20 display the ROC curve for the suggested comparative machine-learning classification models (KLR and SVM). A comparative study with different kernel techniques using the KLR-ROC and SVM-ROC curves across distinct breast disease regions is shown in Fig. 18. Comparative studies of various kernel techniques using the KLR-ROC and SVM-ROC curves across different regions of prostate and colon diseases, respectively, are shown in Figs. 19 and 20.

When the ROC curves of KLR and SVM were compared, it was identified that while the Poly and RBF kernels of KLR were correct in detecting cancerous samples in CDS, the RBF, Poly2, and Linear kernels of SVM were accurate in identifying cancerous samples in breast disease. The Poly2 and Poly3 kernels from SVM showed notable results for cancerous samples in colon disease, whereas the Poly2 and Poly3 kernels from KLR showed excellent results for cancerous samples in colon disease. The Poly2 and Poly3 kernels from KLR were exceptionally effective at identifying cancer in both CDS and N-CDS, while the Linear, Poly2, and RBF kernels from SVM were highly effective at diagnosing prostate cancer in CDS. This examination covered the WR Sequence, N-CDS, and CDS.

The ability to distinguish between non-cancerous and cancerous samples is an essential aspect of extracted features in cancer classification predictions. This study demonstrates the importance of the dinucleotides feature and its interpretability for binary classification. For cancer datasets, a binary classification model used to classify both cancer and non-cancer classes. The SHAP values of the features become important based on the chosen classification approach. A machine learning global interpretable technique known as SHAP (SHapley Additive exPlanations) is used in model predictions to explain the significance of features [40, 41]. It simplifies the evaluation and assessment of models by emphasizing the ways in which each feature is applied to differentiate between samples that are cancerous and non-cancerous. Figures 21–23 display the dinucleotide SHAP values as features for the KLR classification model for CDS, N-CDS, and W.R. sequence regions for breast, colon, and prostate diseases, respectively. The log odds values for both classes are shown on the x-axis in Figs. 21–23, while the feature value for both classes is shown on the y-axis. These features were extracted in order to classify cancer detection in this work, and the features' respective dinucleotide SHAP values show which features have a significant influence on cancer detection.

The KLR and SVM comparison study demonstrated how well the suggested technique worked. As was demonstrated in earlier discussions, the Markov model was able to

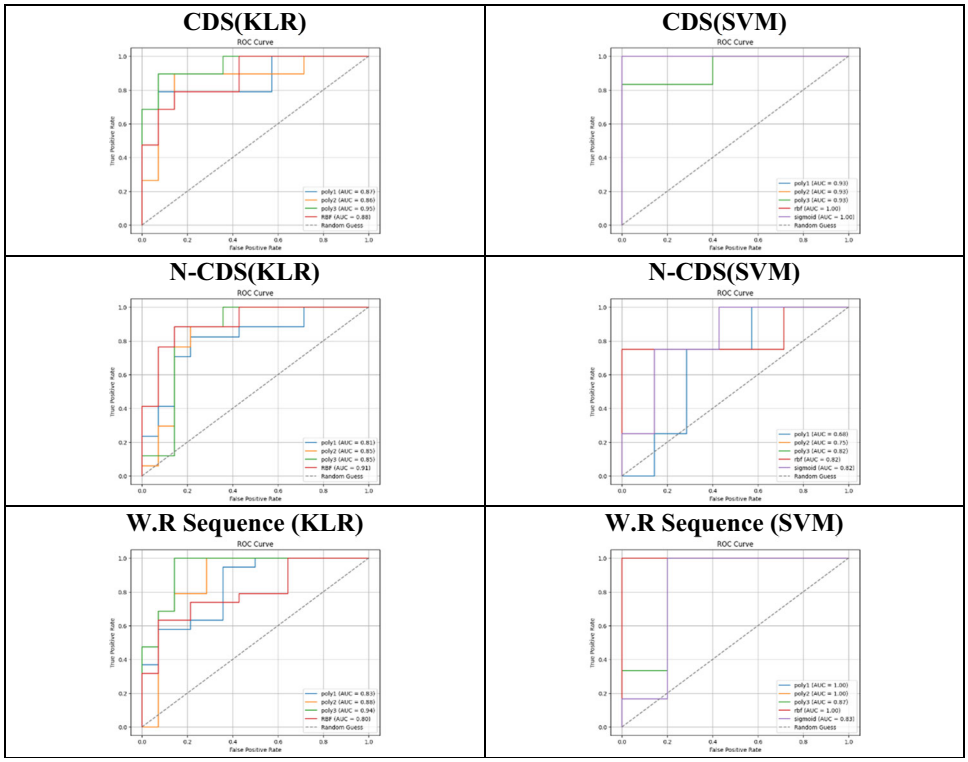

Fig. 19 Comparative ROC Curve Analysis of CDS, N-CDS, and W.R Sequence Regions in Colon Disease

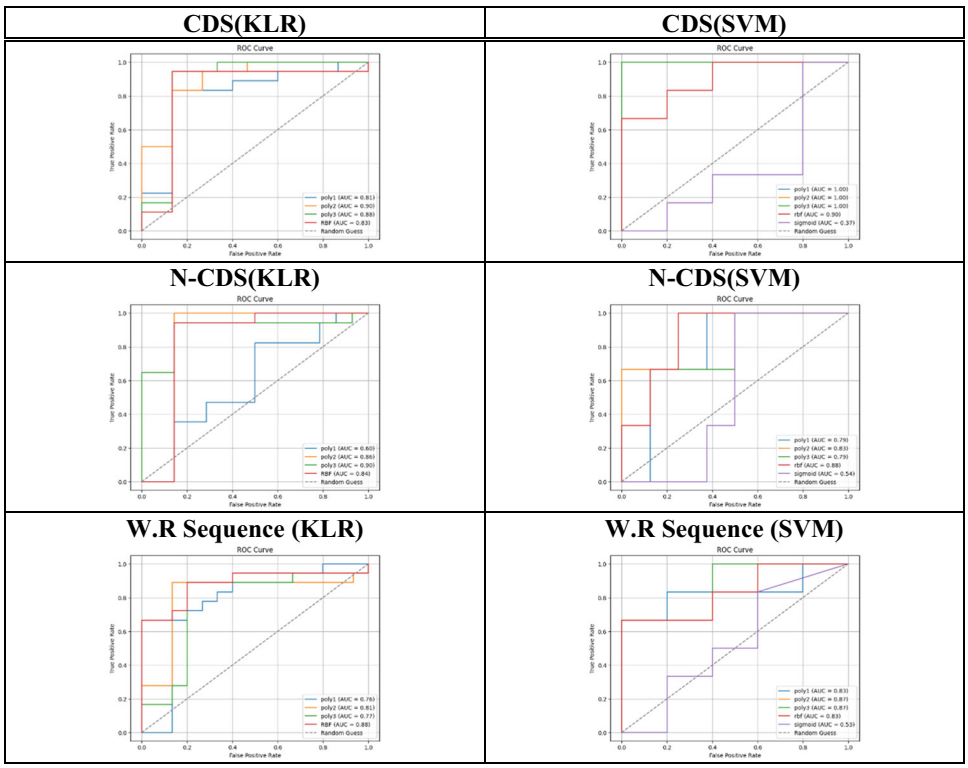

Fig. 20 Comparative ROC Curve Analysis of CDS, N-CDS, and W.R Sequence Regions in Prostate Disease

accurately represent the nucleotide units that make up DNA sequences. The use of the Markov chain to represent nucleotide order in the form of dinucleotide patterns was further explored in this work. When this method was applied to regions, such as CDS, N-CDS, and W.R. sequences, the classification of genomic data showed different successful results. Our study did not focus on any one form of cancer in particular. To identify the gene regions

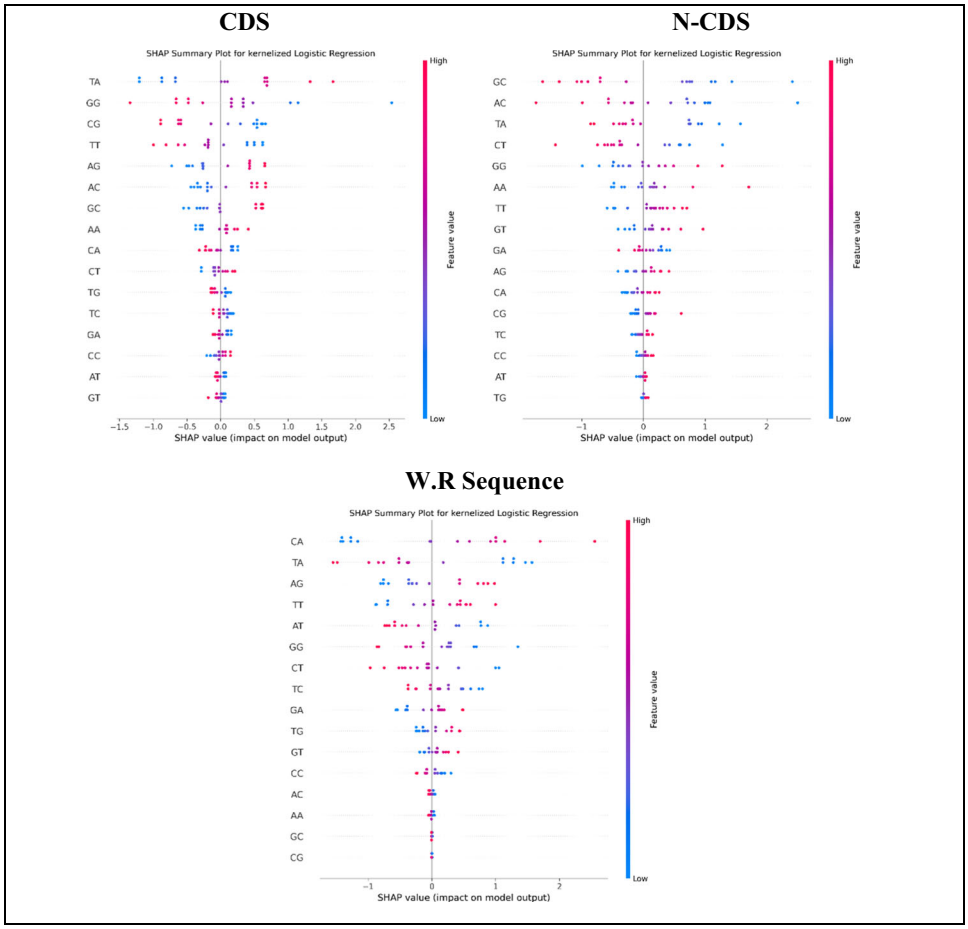

Fig. 21 'SHAP' value of all extracted features for CDS, N-CDS, and W.R Sequence in Breast Disease

related to breast, prostate, and colon diseases, this study compared Kernel Logistic Regression (KLR) with Support Vector Machines (SVM). This helped identify the classification between cases of cancerous and non-cancerous cases. Cancer cases indicate the genetic nature of genetic mutations, in which a few changes in DNA sequences have a major impact on the development of cancer cells. Other genetic diseases are influenced by the development of certain genes that are associated with breast cancer. Since mistakes made at these steps might impact investigations to follow, accurate sequencing and interpretation in the region of DNA sequences are essential.

By using extracted features, the new classification method significantly reduces determining costs when compared to directly classifying high-dimensional DNA sequences. The novel feature dimension reduction approach improves classification accuracy without affecting anything, according to experimental results. The unique advantage of the suggested technique is the presence of relationships at the unit level of DNA sequences. The implementation of a First-order Markov chain-based feature extraction technique allows for this successful outcome. The results highlight a significant relationship between the features that were extracted. These features not only validate the capacity to classify cancerous samples but also offer insights from chemical and genetic points of view. Moreover, modeling for the identification and prediction of cancer-causing genomic sequences can be helped by this technique. The suggested technique also has the advantage of being effective with a smaller set of features, which improves classification performance.

A comprehensive evaluation of several cancer datasets has validated the performance of the implemented methodology. This validates the algorithm's strong application in general. Significantly, the suggested methodology rival's limitations associated with specific cancer types or genes involved. More importantly, our approach stands out due to comprehensive comparisons with both the entire raw DNA sequence and non-coding regions, in contrast to many previous studies that were mainly focused on coding regions [4, 5]. Our research stands out from other studies in this field because of this special feature. Similar to other research, this study takes advantage of the classification phase in addition to the sample feature values [4, 5]. This study is significant because it carefully compares the conditions required to those of previous well-received studies in a related field.

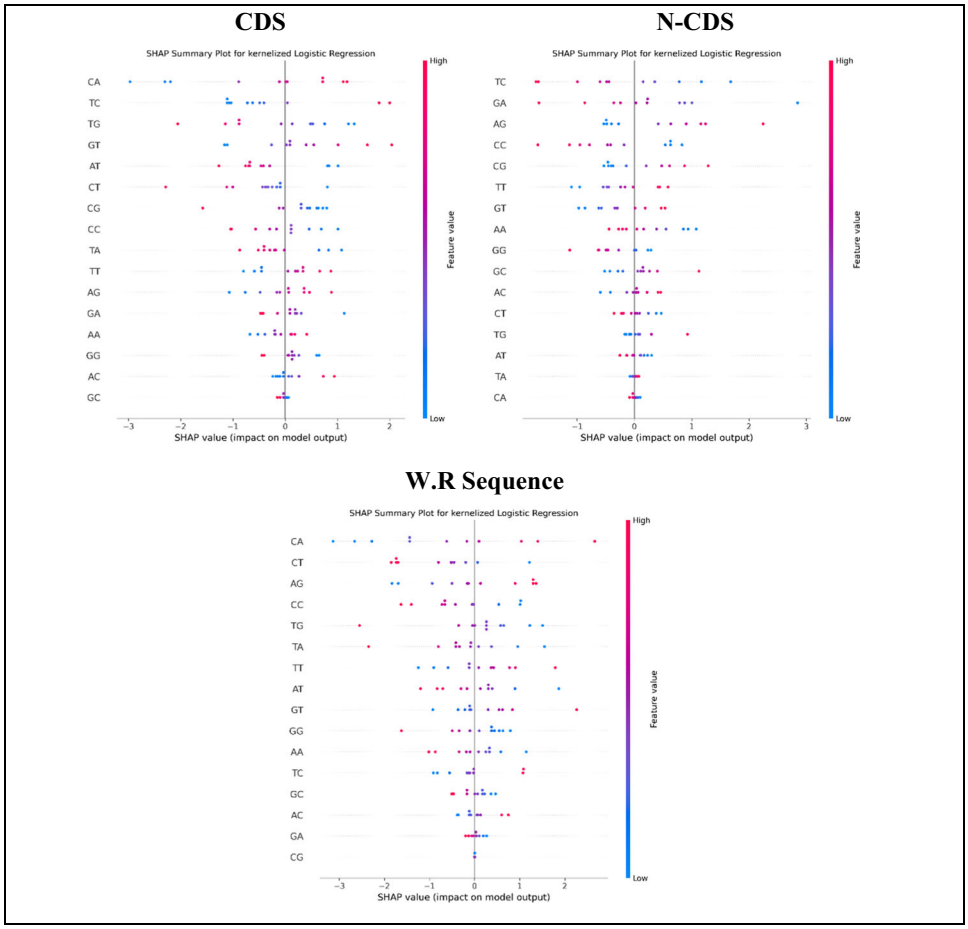

Fig. 22 'SHAP' value of all extracted features for CDS, N-CDS, and W.R Sequence in Colon Disease

Conclusion

Analyzing the complex field of serious diseases like cancer requires in-depth research. Scientists from a wide range of fields are very interested in exploring this field because of its importance. The vast amount of information contained in DNA sequences requires the use of modern feature extraction and selection of data techniques and computational statistical techniques that extend over standard techniques. This study uses computational and statistical approaches to provide its findings in a detailed, point-by-point breakdown as follows:

- The novelty of this work aims to solve problems found in DNA sequence's protein-coding region (CDS), non-protein-coding region (N-CDS), and whole raw DNA sequence including both CDS and N-CDS. In particular, the method of sequential pattern mining as feature extraction and selection of genomes is applied to identify differences and similarities between DNA associated with cancerous and non-cancerous comparable.
- This work presents a unique hybrid method for classifying nucleotide DNA sequences in genes that are cancerous and non-cancerous. The study focuses on the analysis of DNA samples connected to breast, colon, and prostate disease DNA sequences. This approach performs a comparative analysis by combining KLR and SVM techniques through the use of a Markovian feature mapping strategy.
- Our novel feature selection method's initial process takes advantage of this specific observation and performs successfully group-based normalization on features that result from DNA sequences that contain CDS, N-CDS, and Whole Raw DNA Sequences. The results show that a reduced feature, limited to sixteen dimensions, can effectively and significantly discriminate between DNA sequences that are cancerous and non-cancerous.
- According to the simulation results, SVM's RBF, Poly2, and Linear kernels were accurate in breast disease; KLR's Poly and RBF were accurate in CDS. SVM Poly2 and Poly3 indicate significant results in colon disease; KLR's Poly2 and Poly3 showed high levels in CDS. Regarding prostate disease, SVM performed outstandingly in CDS using Linear, Poly2, and RBF; KLR's Poly2 and Poly3 were highly accurate in CDS and N-CDS. Using CDS, N-CDS, and W.R Sequence.

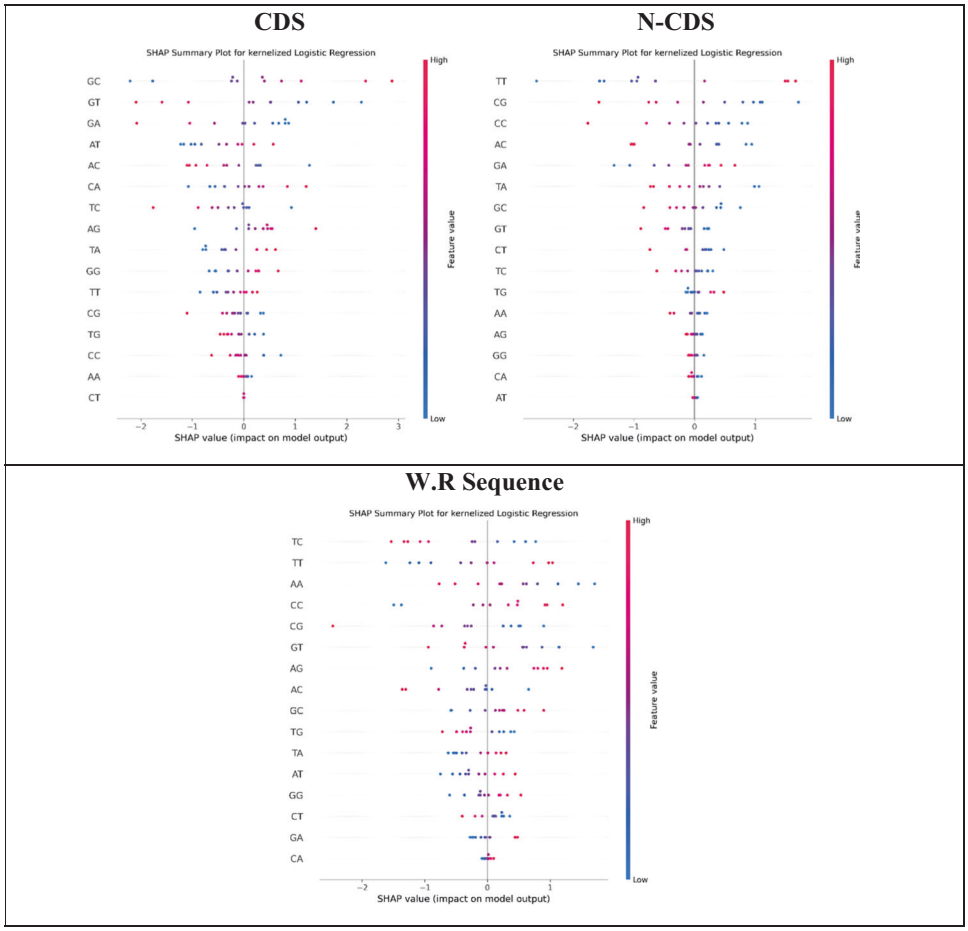

Fig. 23 'SHAP' value of all extracted features for CDS, N-CDS, and W.R Sequence in Prostate Disease

According to the outcomes of our study, our technique has a strong potential for cancer diagnosis by utilizing the most accurate classification models applied to distinct regions of DNA sequences. When analyzing 177 malignant and 161 non-cancerous samples from various cancer types such as breast, colon, and prostate cancer, this novel technique consistently achieves significant accuracy across all detected DNA regions. Notably, our approach is efficient, with a lower computational overhead than other strategies. The ability to analyze vast volumes of DNA sequencing data makes it an appealing alternative for cancer classification.

Future Scope

- Future studies, based on our findings, could examine features that combine the CDS and N-CDS regions, improving DNA sequence classification for cancer identification.
- In the field of cancer detection and classification, distinct statistical techniques are utilized to provide probabilistic features, similar to the Markov model.

Limitation

- Markov models lack long-range DNA patterns that are essential for classifying cancerous from non-cancerous cases.
- In noisy and limited genomic data, estimating transition probabilities in Markov models may be more challenging, which lowers the effectiveness of feature extraction and cancer classification.

Acknowledgements The authors are highly thankful to the education department, Government of Gujarat, India for financial support for carrying out this research work.

Author contributions All authors contributed equally to the design and implementation of the research, to the analysis of the results and to the writing of the manuscript.

Compliance with ethical standards

Conflict of interest The authors declare no competing interests.

References

1. Alberts, B. et al. (2017) *Molecular Biology of the Cell*. https://doi.org/10.1201/9781315735368.
2. Huang, W., Zhang, J., Wang, Y., & Huang, D. (2010) A simple method to analyze the similarity of biological sequences based on the fuzzy theory. *Journal of Theoretical Biology,* 265, 3. https://doi.org/10.1016/j.jtbi.2010.05.008.
3. Khastan, A. & Hooshyar, L. (2019) A computational method to analyze the similarity of biological sequences under uncertainty. *Iranian Journal of Fuzzy Systems*. 16, 6. https://doi.org/10.22111/ijfs.2019.5017.
4. Khodaei, A., Feizi-Derakhshi, M. R., & Mozaffari-Tazehkand B. (2021) A Markov chain-based feature extraction method for classification and identification of cancerous DNA sequences. *BioImpacts*, 11, 2. https://doi.org/10.34172/BI.2021.16.
5. Khodaei, A., Feizi-Derakhshi, M. R., & Mozaffari-Tazehkand, B. (2020) A pattern recognition model to distinguish cancerous DNA sequences via signal processing methods. *Soft Computing*, 24, 21. https://doi.org/10.1007/s00500-020-04942-4.
6. Yang, A., Zhang, W., Wang, J., Yang, K., Han, Y., & Zhang, L. (2020) Review on the application of machine learning algorithms in the sequence data mining of DNA. *Frontiers in Bioengineering and Biotechnology*, 8. https://doi.org/10.3389/fbioe.2020.01032.
7. Sun, Y. et al. (2019) Identification of 12 cancer types through genome deep learning. *Science Reports*, 9, 1. https://doi.org/10.1038/s41598-019-53989-3.
8. Akbar, S., Hayat, M., Iqbal, M., & Jan, M. A. (2017). iACP-GAEnsC: evolutionary genetic algorithm based ensemble classification of anticancer peptides by utilizing hybrid feature space. *Artificial Intellegence in Medicine*, *79*, 62–70. https://doi.org/10.1016/j.artmed.2017.06.008.
9. Akbar, S., Hayat, M., Tahir, M., Khan, S., & Alarfaj, F. K. (2022). cACP-DeepGram: classification of anticancer peptides via deep neural network and skip-gram-based word embedding model. *Artificial Intellegence in Medicine*, *131*, 102349. https://doi.org/10.1016/j.artmed.2022.102349.
10. Akbar, S., Rahman, A. U., Hayat, M., & Sohail, M. (2020). cACP: classifying anticancer peptides using discriminative intelligent model via Chou's 5-step rules and general pseudo components. *Chemometrics and Intelligent Laboratory Systems*, *196*, 103912. https://doi.org/10.1016/j.chemolab.2019.103912.
11. Akbar, S., Hayat, M., Tahir, M., & Chong, K. T. (2020). CACP-2LFS: classification of anticancer peptides using sequential discriminative model of KSAAP and two-level feature selection approach. *IEEE Access*, *8*, 131939–131948. https://doi.org/10.1109/ACCESS.2020.3009125.
12. Pecorino, L. (2012) Molecular Biology of Cancer: Mechanisms, Targets, And Therapeutics. Oxford University Press.
13. Singh, M., Prasad, C. P., Singh, T. D., & Kumar, L. (2018). Cancer research in India: Challenges & opportunities. *Indian Journal of Medical Research*, *148*, 362–365. https://doi.org/10.4103/ijmr.IJMR_1711_18.
14. Zhang, J., Zhang, W., & Yang, H. (2016) In search of coding and non-coding regions of DNA sequences based on balanced estimation of diffusion entropy. *Journal of Biological Physics*, 42. https://doi.org/10.1007/s10867-015-9399-7.
15. Das, J., Barman, S., & Das, J. (2014) Bayesian fusion in cancer gene prediction CODEC design view project genomic signal processing view project Bayesian fusion in cancer gene prediction. CCSN. [Online]. Available: https://www.researchgate.net/publication/280917849
16. Satapathi, G. N., Srihari, P., Jyothi, A., & Lavanya, S. (2013) Prediction of cancer cell using DSP techniques. in *International Conference on Communication and Signal Processing, ICCSP 2013 - Proceedings*. https://doi.org/10.1109/iccsp.2013.6577034.
17. Roy, T., & Barman, S. (2014) A behavioral study of healthy and cancer genes by modeling electrical network. *Gene*, 550. https://doi.org/10.1016/j.gene.2014.08.020.
18. Roy, T., & Barman, S. (2016) Performance analysis of network model to identify healthy and cancerous colon genes. *IEEE Journal of Biomedical and Health Informatics*, 20. https://doi.org/10.1109/JBHI.2015.2408366.
19. Das, J., & Barman, S. (2017) DSP based entropy estimation for identification and classification of Homo sapiens cancer genes. *Microsystem Technologies*, 23 (no. 9). https://doi.org/10.1007/s00542-016-3056-3.
20. Singha Roy, S., & Barman, S. (2021) A non-invasive cancer gene detection technique using FLANN based adaptive filter. *Microsystem Technologies*, 27 (no. 2). https://doi.org/10.1007/s00542-018-4036-6.
21. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015) Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13. 2015. https://doi.org/10.1016/j.csbj.2014.11.005.
22. Margaliot, M. (2008) Pattern Recognition (Theodoridis, S. and Koutroumbas, K.; 2006) [Book reviews]. *IEEE Transactions on Neural Networks*, 19 (no. 2). https://doi.org/10.1109/tnn.2008.929642.
23. SenthilVelMurugan, N., Vallinayagam, V. V. V., Senthamarai Kannan, & Viveka, T. (2013) Analysis of liver cancer DNA sequence data using data mining. *International Journal of Computer Application*, 61 (no. 3). https://doi.org/10.5120/9909-4502.
24. Blitzstein, J. K., & Hwang, J. (2014) *Introduction to probability*. https://doi.org/10.1201/b17221.
25. Fernandes, A. A. T., Filho, D. B. F., da Rocha, E. C., & da Silva Nascimento, W. (2020) Read this paper if you want to learn logistic regression. *Revista de Sociologia e Politica*, vol. 28 (no. 74). https://doi.org/10.1590/1678-987320287406EN.
26. Liu, L. (2018) Research on logistic regression algorithm of breast cancer diagnose data by machine learning. in *Proceedings - 2018 International Conference on Robots and Intelligent System, ICRIS 2018*, Institute of Electrical and Electronics Engineers Inc., pp. 157–160. https://doi.org/10.1109/ICRIS.2018.00049.
27. Ha, J., Kambe, M., & Pe, J. (2011) Data Mining, *Data Mining: Concepts and Techniques*. https://doi.org/10.1016/C2009-0-61819-5.
28. Dong, G., & Pei, J. (2007) Classification, clustering, features and distances of sequence data. in *Sequence Data Mining*, 47–65. https://doi.org/10.1007/978-0-387-69937-0_3.
29. Shaikh, F. J., & Rao, D. S. (2021). Prediction of cancer disease using machine learning approach. *in Materials Today: Proceedings*, *50*, 40–47. https://doi.org/10.1016/j.matpr.2021.03.625.
30. De Ridder, D., De Ridder, J., & Reinders, M. J. T. (2013) Pattern recognition in bioinformatics. *Briefings in Bioinformatic*, 14 (no. 5). https://doi.org/10.1093/bib/bbt020.
31. Rong, M. L. K., Kuruoglu, E. E., & Chan, W. K. V. (2023) Modeling SARS-CoV-2 nucleotide mutations as a stochastic process. *PLoS One*, 18 (no. 4). https://doi.org/10.1371/journal.pone.0284874.
32. Rymarczyk, T., Kozłowski, E., Kłosowski, G., & Niderla, K. (2019) Logistic regression for machine learning in process tomography. *Sensors (Switzerland)*, 19 (no. 15). https://doi.org/10.3390/s19153400.

33. Burge, C. B., & Karlin, S. (1998) Finding the genes in genomic DNA. *Current Opinion in Structural Biology*, 8 (no. 3). https://doi.org/10.1016/S0959-440X(98)80069-9.
34. GenBank National Center for Biotechnology Information Database. Available from: http://www.ncbi.nlm.nih.gov.
35. Pham, B. T. et al. (2020) A comparative study of kernel logistic regression, radial basis function classifier, multinomial naive bayes, and logistic model tree for flash flood susceptibility mapping. *Water (Switzerland)*, 12 (no. 1). https://doi.org/10.3390/w12010239.
36. Cawley, G. C., & Talbot, N. L. C. (2008) Efficient approximate leave-one-out cross-validation for kernel logistic regression. *Machine Learning*, 71 (no. 2–3). https://doi.org/10.1007/s10994-008-5055-9.
37. Tien Bui, D., Tuan, T. A., Klempe, H., Pradhan, B., & Revhaug, I. (2016) Spatial prediction models for shallow landslide hazards: a comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree. *Landslides*, 13 (no. 2). https://doi.org/10.1007/s10346-015-0557-6.
38. Cawley, G. C., & Talbot, N. L. C. (2004). Efficient model selection for kernel logistic regression. *in Proceedings - International Conference on Pattern Recognition*, *2*, 439–442. https://doi.org/10.1109/ICPR.2004.1334249.
39. Breneman, J. (2005) Kernel methods for pattern analysis. *Technometrics*, 47 (no. 2). https://doi.org/10.1198/tech.2005.s264.
40. Amami, R., Ben Ayed, D., & Ellouze, N. (2012). An empirical comparison of SVM and some supervised learning algorithms for vowel recognition. *International Journal of Intelligent Information Processing*, *3*(no. 1), 63–70. https://doi.org/10.4156/ijiip.vol3.issue1.6.
41. Raza, A., Uddin, J., Almuhaimeed, A., Akbar, S., Zou, Q., & Ahmad, A. (2023) AIPs-SnTCN: predicting anti-inflammatory peptides using fasttext and transformer encoder-based hybrid word embedding with self-normalized temporal convolutional networks. *Journal of Chemical Information and Modelling*. 63 (no. 21). https://doi.org/10.1021/acs.jcim.3c01563.
42. Akbar, S., Zou, Q., Raza, A., & Alarfaj, F. K. (2024). iAFPs-Mv-BiTCN: Predicting antifungal peptides using self-attention transformer embedding and transform evolutionary based multi-view features with bidirectional temporal convolutional networks. *Artif Intell Med*, *151*, 102860 https://doi.org/10.1016/j.artmed.2024.102860. p. 102860, May.