

## A novel method for examining promoters using statistical analysis and artificial intelligence learning

Sinan Salim Mohammed Sheet<sup>1</sup>, Marwa Mawfaq Mohamedsheet Al-Hatab<sup>2</sup>, Maysaloon Abed Qasim<sup>3</sup>

<sup>1</sup>Technical Medical Instrumentation, Polytechnic College Mosul, Northern Technical University, Mosul, Iraq

<sup>2</sup>Technical Engineering College, Northern Technical University, Mosul, Iraq

<sup>3</sup>Technical Engineering College for Computer and Artificial Intelligence, Northern Technical University, Mosul, Iraq

### Article Info

#### Article history:

Received Nov 9, 2024

Revised Jul 15, 2025

Accepted Aug 6, 2025

#### Keywords:

Area under the curve

Deoxyribonucleic acid

Machine learning

Promoter

Statistical feature analysis

### ABSTRACT

Accurately classifying promoters has become a significant focus in bioinformatics research. Although numerous studies have attempted to address this challenge, the performance of existing methods still leaves room for improvement. This study, statistical feature analysis has been applied to the features that have been developed in our previous work. This approach extracted additional informative features from basic sequence characteristics and then used them together with the original and newly engineered features. Utilizing statistical feature analysis enhanced key patterns, which lead to an improvement in the accuracy of the promoter classification. Results demonstrated that our proposed method outperforms other models that use only basic features. The value of the area under the curve (AUC) of 0.83958 achieved when using the combined feature set confirmed the effectiveness of our approach. Furthermore, the AUC value reached 1 when these optimized features were used with naive Bayes (NB) classifier, referring to the strength of incorporating statistical analysis into feature design.

This is an open access article under the [CC BY-SA](#) license.



### Corresponding Author:

Sinan Salim Mohammed Sheet

Technical Medical Instrumentation, Polytechnic College Mosul, Northern Technical University

Mosul, Iraq

Email: sinan\_sm76@ntu.edu.iq

## 1. INTRODUCTION

Regulation of gene expression is a vital cellular process that ensures development, physiological balance, and adaptation to environmental changes. It determines when and how genes are expressed, shaping protein diversity and cellular identity [1]. Dysregulation of this process is closely linked to human diseases such as cancer, metabolic disorders, and neurological conditions. Promoter regions which are short deoxyribonucleic acid (DNA) stretches upstream of genes that act as control hubs for transcription initiation are among the critical regulators [2].

Promoter regions provide docking sites for ribonucleic acid (RNA) polymerase and transcription factors. Early studies described essential motifs like the -35 (TTGACA) and -10 (TATAAT) elements in bacterial promoters, with transcription starting near a purine downstream of the -10 box. However, promoter structures vary widely across species [3].

Identifying promoters remains challenging because many lack conserved motifs and overlap with other regulatory regions. The accurate detection is very complicated due to their sequence variability, chromatin structure, and species-specific differences. Traditional computational methods, relying on motifs or position weight matrices, often suffer from low accuracy and high false discovery rates, limiting their reliability for large-scale genomic studies [4].

To overcome these challenges, researchers have turned to artificial intelligence (AI) based approaches. Machine learning (ML) and deep learning (DL) models can capture both sequence level motifs and long-range dependencies, improving prediction performance. Methods such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and hybrid CNN-long short-term memory (LSTM) models have achieved good results, while attention mechanisms and Transformer-based architectures offer new possibilities for modeling promoter complexity [5]–[7].

Despite these advances, AI based methods still face problems of data scarcity and interpretability. Large, high-quality datasets are often unavailable. This study aims to address these gaps by developing robust AI-based promoter identification methods that integrate biological knowledge with ML. The next section reviews previous methods, with emphasis on the evolution from traditional models to modern AI-driven approaches.

## 2. LITERATURE REVIEW

Because the methods which depend on traditional laboratory are often resource-intensive, slow, and not scalable for whole-genome studies, computational approaches have become a very important tools in the prediction of promoto [8]. ML techniques among these computational approaches have been emerged due to their particularly effective, capable of uncovering intricate sequence patterns and dependencies that might be overlooked by conventional algorithms. ML models can accurately distinguish between promoter regions and background genomic sequences with high predictive power by converting raw DNA sequences into structured feature representations.

Amin *et al.* [9] proposed a study using a DL-based approach in identification and classification of bacterial sigma promoters using branched CNNs. Their method which is called prompt-learning pre-trained language model for promoter prediction (PLPMpro), has been designed to distinguish between promoter and non-promoter sequences in addition to promoters' classification into different sigma factor categories, such as  $\sigma^{70}$  and  $\sigma^{32}$ . The system used parallel convolutional branches to extract diverse feature representations from DNA sequences, and this result in an improvement in the classification performance. Their CNN-based framework achieved a accuracy and generalizability in both binary and multiclass promoter prediction tasks [9].

Tayara *et al.* [10] in the same year, introduced a hybrid deep-learning framework called identification of prokaryotic promoters and their strength via windows (iPSW) using pseudo dinucleotide composition (PseDNC)-based deep learning to be used in the identification of prokaryotic promoters and classify them into two categories, strong and weak. The study integrates between CNNs and PseDNC. This hybrid architecture has been applied on benchmark Ecoli datasets and showed high accuracy in promoter detection [10].

Moraes *et al.* [11] proposed CapsProm, which is a capsule network-based model used to identify promoter across seven different organisms, including eukaryotes and prokaryotes. CapsProm get benifit from the ability of the capsule network' to maintain hierarchical relationships within sequence patterns. This method demonstrated competitive F1-scores surpassing baseline CNN approaches in five out of seven datasets. The authors emphasized the generalizabilityof the CapsProm's gsystem, according to its strength in cross-species promoter prediction and potential for transfer learning (TL) across genomic contexts [11].

Zhang *et al.* [12] introduced a model for promoter prediction. This model produces a hybrid DL framework combining CNNs, capsule networks, bidirectional long short-term memory (Bi-LSTM), and a self-attention mechanism to identify promoters effectively and classify their strength. It uses one-hot encoding to represent DNA sequences and gets benifits from both local and global sequence features to enhance prediction performance. The model has achieved an accuracy of approximately 86% for promoter identification and around 73.5% for promoter strength classification [12].

In another related study, Li *et al.* [13] developed a novel approch PLPMpro. This approach enhanced the prediction of the promotor sequence by combining the prompt-learning with pre-trained language models. Their study used prompt-based fine-tuning to leverage genomic representations learned from large-scale training corpora, which increase the ability of the system to capture complex promoter sequence features more effectively. After evaluated the system on benchmark datasets from the Eukaryotic promoter database, the results achieved in both precision and recall demonstrated notable improvements comparing to conventional transformer-based models such as DNA bidirectional encoder representations from transformers (BERT) [13].

Paul *et al.* [14] developed machine learning and duplex stability promoter prediction (MLDSPP) named system focusing on bacterial genomes. This study is a tool designed to detect promotor regions cross 12 prokaryotic species. This method used ML algorithms such as extreme gradient boosting (XGBoost) with structural DNA features such as duplex stability. The results obtained from using MLDSPP demonstrated a superiority to existing tools like Sigma70pred and iPromoter2L, which achieved F1-scores above than 95%.

Moreover, the study used explainable AI techniques, including Shapley values and one-hot encoding, to improve the transparency of the model and increase the predictive accuracy [14].

Ashayeri *et al.* [15] applied TL techniques on several genomic tasks, such as analysis of gene expression, detect of mutation, and recognition of genetic syndrome. Results showed that by using TL, the efficiency and accuracy of the model has been significantly improved in various genetic research domains. In addition TL enhances the accuracy and efficiency of mutation detection, which can help in identifying genetic abnormalities, and it is also able to improve diagnostic accuracy of syndrome-related genetic patterns. Furthermore, TL contributes in gene expression analysis by enabling more precise predictions of expression levels and their relationships. It also can strength the studies related to phenotype-genotype by using knowledge from pre-trained models [15].

Zeng *et al.* [16] introduces a novel DNA sequence segmentation method and a refined dictionary for BERT pre-training, enhancing promoter detection through DL techniques like CNNs, LSTMs, and Inception networks, improving performance and interpretability in downstream tasks [16]. Finally, Gunarathna *et al.* [17] employed interpretable ML models guided by assay for transposase-accessible chromatin using sequencing (ATAC-seq) data to uncover cancer-specific chromatin features in cell-free deoxyribonucleic acid (cfDNA). Their approach focused on enhancing the prediction of breast cancer-derived cfDNA by leveraging from the chromatin accessibility signals, which have led to improved detection performance. Although their findings highlighted the potential of chromatin-based features in non-invasive cancer diagnostics, the study did not directly address promoter region identification [17].

While several ML-based promoter detection methods exist, many rely on generic features or limited nucleotide compositions, often resulting in low accuracy. This study addresses this limitation by introducing novel statistical and biological features specifically designed for promoter detection. The main objective is to assess the effectiveness of these features in improving ML classifier performance. To this end, we employed support vector machine (SVM), logistic regression (LR), k-nearest neighbors (KNN), decision tree (DT), and naive Bayes (NB). These classifiers were selected for their complementary strengths: i) SVM handles high-dimensional and non-linear data, ii) LR offers interpretable linear modeling, iii) KNN captures local sequence similarities, iv) DT effectively manages feature interactions, and v) NB performs well under probabilistic assumptions. This diverse classifier selection ensures a comprehensive evaluation of the proposed features.

### 3. METHODOLOGY

This section illustrates the overall methodology used in this study. It starts from data preprocessing and feature extraction methods to the model development and performance evaluation. Figure 1 shows the workflow in this study.

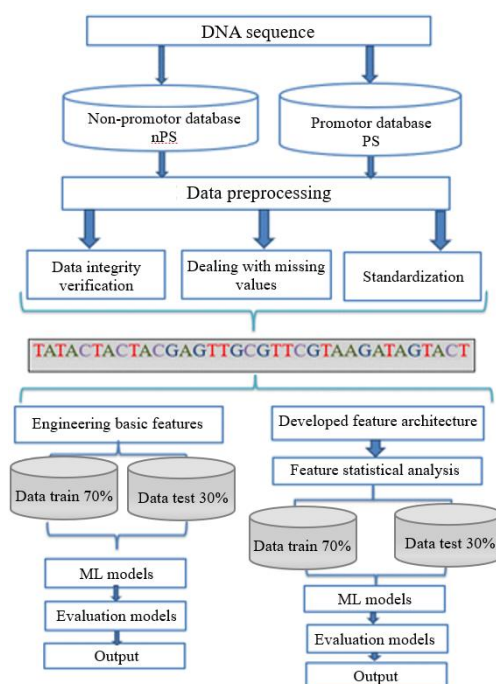


Figure 1. Flowchart of the proposed model

### 3.1. Data preprocessing

The dataset used in this study has been obtained from the University of California Irvine (UCI) ML repository [18], it consists of 106 nucleotide sequences, each sequence length has 57 base pairs, spanning positions -50 to +7. These sequences have been divided into two categories: promoter sequence as positive class (PS) and non-promoter sequence as negative class (n\_PS). The preprocessing of the data begins with splitting the dataset into training and testing subsets, then the data has been checked and corrected to ensure sequence accuracy and completeness. Depending on the extent of missing information, missing or unclear identifiers within both classes were addressed through either imputation or removal. Last operation in the preprocessing was the cleaning and normalization of the sequences by eliminating extraneous elements standardize their format and maintain consistency for subsequent analyses.

### 3.2. Feature engineering

#### 3.2.1. Basic feature engineering

Basic feature engineering method analyses DNA sequences based on the components of their fundamental nucleotide-adenine (A), thymine (T), cytosine (C), and guanine (G). Each DNA sequence was broken down into individual nucleotides, and each nucleotide referred to as a separate feature. This method can identify short, localized nucleotide patterns which are important in distinguishing between PS and n\_PS types.

#### 3.2.2. Developed feature engineering

The aim of developed feature engineering approach is to enhance the accuracy of classification by extracting a comprehensive set of biologically meaningful attributes from DNA sequences. This method integrates different evaluation in order to capture both global and local sequence characteristics, these evaluations are nucleotide composition analysis, GC content measurement, k-mer frequency profiling, and sequence complexity evaluation. Nucleotide counting determine the occurrences of adenine, thymine, cytosine, and guanine. GC content analysis measures frequency of guanine and cytosine nucleotides which is important in DNA stability according to their triple hydrogen bonds. K-mer analysis investigates recurring nucleotide motifs of length finally, sequence complexity analysis assesses the variability and irregularity in nucleotide distribution. Table 1 illustrates the significant compositional and structural differences between PS and n\_PS by using Developed Features.

### 3.3. Feature statistics and significance biological performance metrics

Different statistical and evaluation metrics have been used, in order to assess the significance and performance of each feature in the classification task. These metrics give accurate analysis for feature distributions and their relationships with the classification results. The metrics used in this study were correlation coefficients, root mean square error (RMSE), mean and standard deviation (SD), signal-to-noise ratio (SNR), and the area under the curve (AUC). The formulas for these metrics are detailed as follows [19], [20].

- i) Correlation: the correlation coefficient measuring the relationship between each feature  $x$  and the classification target  $y$ . In (1) shows the mathematical formula of correlation:

$$\text{correlation}(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (1)$$

where  $x_i, y_i$  are the values of feature and target for sample  $i$ ,  $\bar{y}, \bar{x}$  are the means of  $X$  and  $Y$ ,  $n$  is the number of samples.

- ii) Root mean square (RMS): RMS is used to assess the average magnitude of a feature as (2).

$$\text{RMS}(x) = \sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2} \quad (2)$$

- iii) Mean and SD: the mean and SD of a feature describe its central tendency and variability is shown in (3) and (4):

$$\text{mean}(x) = \frac{1}{n} \sum_{i=1}^n x_i \quad (3)$$

$$\text{STD}(x) = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \quad (4)$$

- iv) SNR: SNR quantifies how much signal is present in a feature relative to its noise as in (5):

$$SNR(x) = \frac{\bar{x}}{STD(x)} \quad (5)$$

- v) AUC: to determine the contribution of each feature in the performance of the model, an ablation study was performed. Each feature has been removed individually, and the classifier has been retrained and the AUC of the model without this feature was recorded. In (6) used to determine AUC difference due to removal of feature  $x_j$  is:

$$\Delta AUC_j = AUC_{full} - AUC_{-x_j} \quad (6)$$

where,  $AUC_{full}$  is the model performance with all features,  $AUC_{-x_j}$  is the performance after removing feature  $x_j$ .

Table 1. Summary of developed features for PS and n\_PS

Feature	Nucleotide	PS average value	n_PS average value	Biological significance
Nucleotide count	Adenine (A)	15.79	14.02	A appears more often in PS regions, playing a role in facilitating DNA strand separation and initiating transcription.
	Thymine (T)	17.19	15.11	A high presence of T in PS regions enhances DNA flexibility, making it easier to unwind the strands during transcription.
	Cytosine (C)	12.62	13.51	A low count of C content in PS regions results in diminished structural stability of the DNA.
	Guanine (G)	11.4	14.45	A decreased level of G in PS regions enhances accessibility for the transcription machinery.
Nucleotide count range (per 57 nucleotides)	Adenine (A)	15–18	13–15	In PS regions, high A content aids in DNA unwinding, whereas n_PS regions display a more balanced nucleotide composition.
	Thymine (T)	16–19	14–16	Increased T levels in PS regions contribute to greater DNA flexibility, while n_PS regions preserve structural stability.
	Cytosine (C)	11–13	13–14	A decline in C content within PS areas leads to reduced DNA stability, facilitating transcription.
	Guanine (G)	10–12	14–15	Less G in PS regions improves access for transcription factors.
GC content (%)	—	40–45%	48–52%	A lower GC content in PS enhances DNA flexibility, whereas higher GC content in n_PS strengthens DNA structure.
K-mer analysis	—	Common motifs such as TATA, CGG, and GCG occur frequently, indicating a rich presence of regulatory sequences	Irregular or loosely organized patterns with no recurring motifs	Specific, organized motifs in PS regions help control gene expression; such motifs are generally absent in n_PS regions.
Sequence complexity	—	Elevated complexity with diverse motifs and structural elements	Limited complexity, characterized by basic and repetitive sequences	The greater sequence complexity found in PS refers to the presence of regulatory elements, while the lower complexity in n_PS implies minimal regulatory function.

### 3.4. Classifier initialization and model selection

#### 3.4.1. Support vector machine

SVM is an effective classifier for handling complex, high-dimensional data by maximizing the margin between classes using kernel functions [21]. A linear kernel was determined using (7) [22]:

$$f_{linear}(x) = \sum_{i=1}^n \alpha_i y_i(x \cdot x_i) + b \quad (7)$$

where  $\alpha_i$  is the Lagrange multiplier,  $y_i$  class labels, and  $x_i$  support vectors.

### 3.4.2. K-nearest neighbors

KNN is a non-parametric, instance-based learning algorithm that classifies a sample based on the majority label among its k closest neighbors in the feature space [23], as in (8):

$$U = \arg \max_U \sum_{i=1}^k I(U_i = U) \quad (8)$$

where  $I(U_i = U)$  represent the indicator function, if  $(U_i = U)$  the value is 1 and otherwise 0. k is several nearest neighbors.

### 3.4.3. Logistic regression

LR is a widely-used linear model that estimates the probability of class membership through a logistic function. Its simplicity allows for straightforward interpretation of feature contributions via model coefficients [24]. The mathematical formula shown in (9):

$$P(y = 1|X) = \frac{1}{1 + e^{-(w \cdot X + b)}} \quad (9)$$

where  $X$  is feature vector,  $w$  represents the weight vector, and  $b$  is the bias term.

### 3.4.4. Naive Bayes

NB classifiers rely on strong conditional independence assumptions between features to compute posterior probabilities efficiently. Despite its simplicity, NB performs surprisingly well in high-dimensional spaces and is particularly effective when the dataset meets or approximates these probabilistic assumptions. Its fast training and inference times make NB a useful benchmark for probabilistic classification models [25]:

$$P(y|X) = \frac{P(y) \prod_{i=1}^n P(x_i|y)}{P(X)} \quad (10)$$

where the prior probability of the class is represented by  $P(y)$ ,  $P(x_i|y)$  is the probability of a feature and  $x_i$  is the given class.

### 3.4.5. Decision tree

DT classify data by recursively splitting the feature space based on thresholds that maximize class separation [26]. In (12) shows the mathematical formula:

$$Gini(t) = 1 - \sum_{i=1}^c p(i|t)^2 \quad (11)$$

where  $p(i|t)$  is the proportion of class  $i$  at node  $t$ .

## 3.5. Performance evaluation

In this study, different metrics have been used to evaluate classification of each ML model [27], [28].

- i) Accuracy: this metric represents the ratio of correctly classified samples to the total number of samples [20]. It is calculated as shown in (12):

$$A = \frac{True\ Positive\ (TP) + True\ Negative\ (TN)}{True\ Positive\ (TP) + True\ Negative\ (TN) + False\ Positive\ (FP) + False\ Negative\ (FN)}$$

where,  $A$  denotes accuracy, TP and TN are the correctly predicted positive and negative cases, respectively, while FP and FN represent false predicted positive and negative cases.

- ii) Precision: precision is the ratio of TP predictions to all positives predicted, as in (12).

$$Precision = \frac{TP}{TP + FP} \quad (12)$$

- iii) Recall (sensitivity): in (13) represents the recall (sensitivity) and indicates actual positives.

$$Recall = \frac{TP}{TP + FN} \quad (13)$$

- iv) F1-score: F1- score refers to actual positives, (14) shows the mathematical formula to determine the F1-score:

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (14)$$

## 4. RESULTS AND DISCUSSION

### 4.1. Biological features evaluation

Table 2 summarizes key statistical and performance metrics for the features, including correlation with the target, RMS, mean, STD, SNR, and AUC from the ablation study. Table 2 together with Figure 2 also highlight feature relevance. Basic nucleotide counts (Count\_A, Count\_T, Count\_C, Count\_G) and GC\_Content show the strongest predictive power, with Count\_T and Count\_A positively correlated and Count\_G and GC\_Content negatively correlated with classification. Sequence\_Complexity, despite a moderate AUC (0.7380), has a high SNR (~42), indicating stable, valuable input. Sequence\_Variability has low SNR and correlation, suggesting limited standalone usefulness but possible value when combined.

Table 2 .Summary of feature statistics and performance metrics from ablation study

Feature	Correlation	RMS	Mean	STD	SNR	AUC
Count_C	-0.14549	13.417	13.066	3.0621	4.267	0.7682
Count_A	0.25438	15.305	14.858	3.689	4.028	0.7419
Count_T	0.26664	16.613	16.151	3.9104	4.13	0.7508
Count_G	-0.43651	13.39	12.925	3.5178	3.674	0.739
GC_Content	-0.43668	0.4628	0.45597	0.07959	5.729	0.754
Sequence_Complexity	-0.40924	1.9487	1.9482	0.0462	42.162	0.738
Sequence_Variability	0.40644	0.0078	0.0059	0.0052	1.134	0.7378

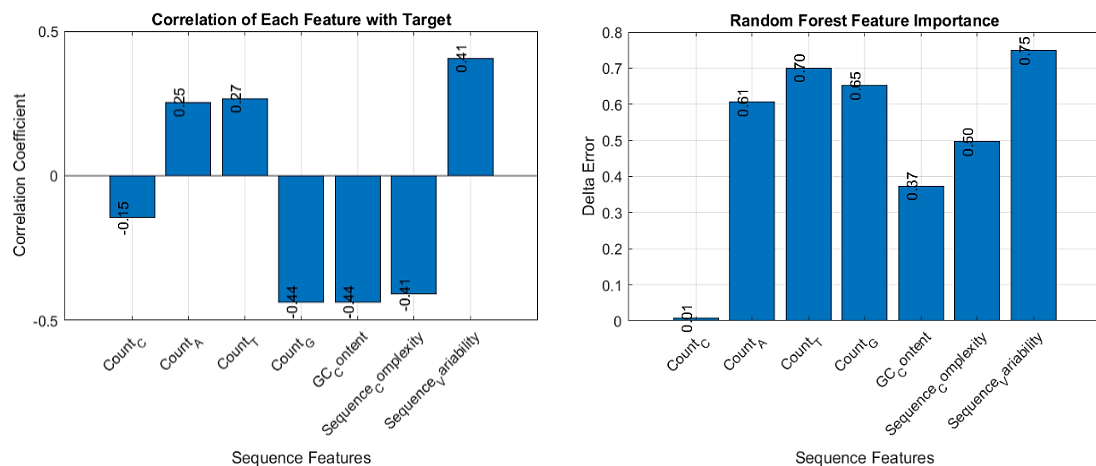


Figure 2. Feature importance and predictive value based on AUC and statistical stability

### 4.2. K-mer pattern analysis

K-mer analysis was performed to link short nucleotide motifs with promoter classification. Each sequence was labeled and annotated with its top three frequent 3-mers, which were broken down into k-mers to calculate class-specific frequencies. Statistical tests (Chi-square or Fisher's exact) assessed k-mer significance across classes. While some k-mers appeared class-specific (e.g., 'aac, acg, cgc' in Class 0; 'aaa, ata, taa' in Class 1), most tests showed non-significant results, likely due to small sample size and sparse data (e.g., Chi-square  $p=1.0000$ ). These results suggest k-mers alone have limited discriminative power but can enhance models when combined with other features, as illustrated in Figure 3.

### 4.3. Classifiers for engineering features

Table 3 and Figure 4 show the performance of different classifiers using basic features. SVM achieved 65% accuracy but had low specificity (0.56) and moderate precision (0.61) despite good sensitivity (0.73). KNN performed poorly with 48% accuracy and very low specificity (0.25), struggling to classify  $n_{PS}$  correctly. LR showed balanced results with 61% accuracy and 0.5 specificity. DTs performed better, reaching 71% accuracy and 0.69 specificity and precision. NB was the best, achieving 90% accuracy, 0.94 specificity, and 0.93 precision.

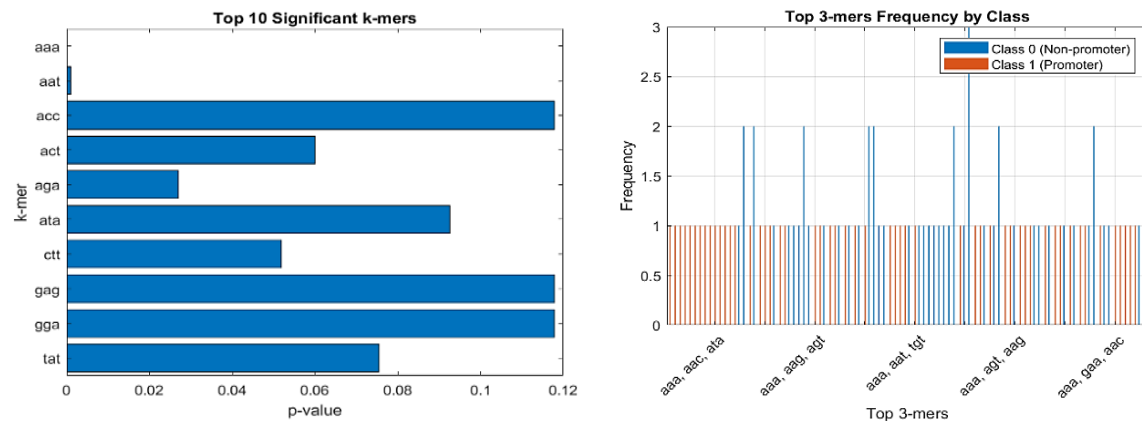


Figure 3. K-mer frequency patterns

Table 3. The performance metrics using basic features

Models	Accuracy (%)	Precision	F1-score	Sensitivity	Specificity
SVM	0.56	0.73	0.67	0.61	0.65
KNN	0.25	0.73	0.58	0.48	0.48
LR	0.5	0.73	0.65	0.58	0.61
DT	0.69	0.73	0.71	0.69	0.71
NB	0.87	0.9	0.93	0.9	0.94

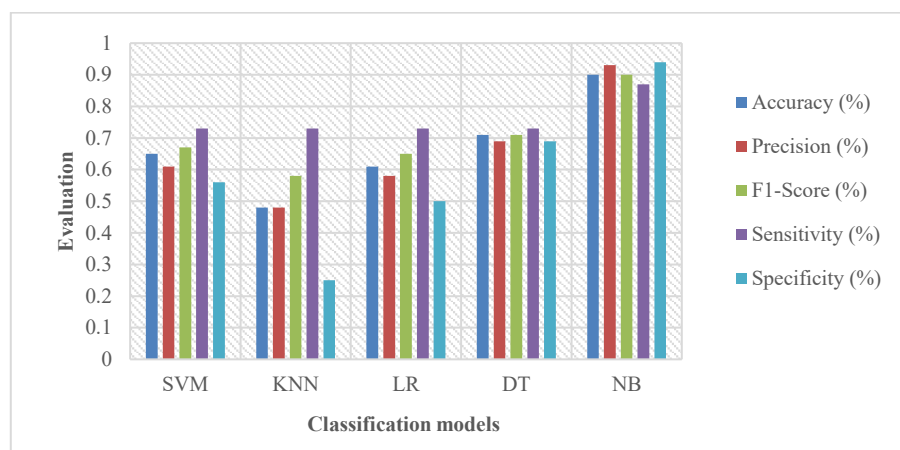


Figure 4. The performance metrics of classifiers for engineering basic features

Table 4 and Figure 5 present results using the newly developed features, demonstrating significant improvement across classifiers, especially for those that struggled with basic features. Enhanced features incorporating domain knowledge and higher-order sequence information helped SVM and KNN better capture non-linear patterns, improving accuracy and specificity. DTs and LR also showed gains in recall, precision, and F1-score. Overall, the new feature set boosted all classifiers, with SVM and KNN becoming far more competitive, reflecting the clear advantage of the proposed feature engineering over traditional methods.

Table 4. The performance metrics using enhanced feature architecture

	Accuracy (%)	Precision	F1-Score	Sensitivity	Specificity
SVM	0.75	0.87	0.81	0.76	0.81
KNN	0.63	0.8	0.73	0.67	0.71
LR	0.63	0.87	0.76	0.68	0.74
DT	0.81	0.87	0.84	0.81	0.84
NB	1	1	1	1	1



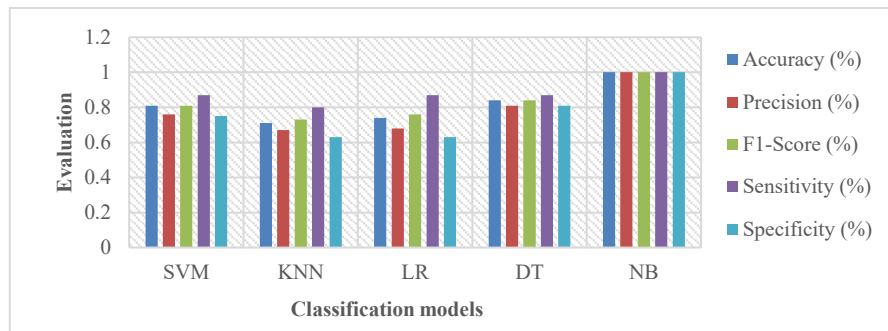


Figure 5. The performance metrics of classifiers for developed feature architecture

Figure 6 compares receiver operating characteristic (ROC) curves and metrics for classification using conventional features Figure 6(a) versus the proposed enhanced features Figure 6(b). The improved feature set clearly boosts most classifiers' performance. DT achieve a solid AUC of 0.83958, while NB reaches a perfect 1.0 with the enhanced features, showing excellent discrimination between PS and n\_PS. KNN struggles the most, with the lowest AUC of 0.7125, and SVM and LR perform only moderately (AUCs of 0.80833 and 0.74583, respectively). NB remains strong with an AUC of 0.90208 even using conventional features. KNN's poor performance across both feature sets likely stems from its sensitivity to feature dimensionality and complexity. These results highlight that the improved features provide a more robust data representation.

However, this study is not without limitations. The dataset is limited in size and variety, which may affect the ability of the model to generalize to broader biological contexts. Also, the features focus mainly on nucleotide composition, ignoring important biological factors like transcription factor binding sites or epigenetic modifications. Future research must address these gaps by expanding the dataset, integrating richer biological data, and exploring advanced DL techniques to achieve better predictive accuracy.

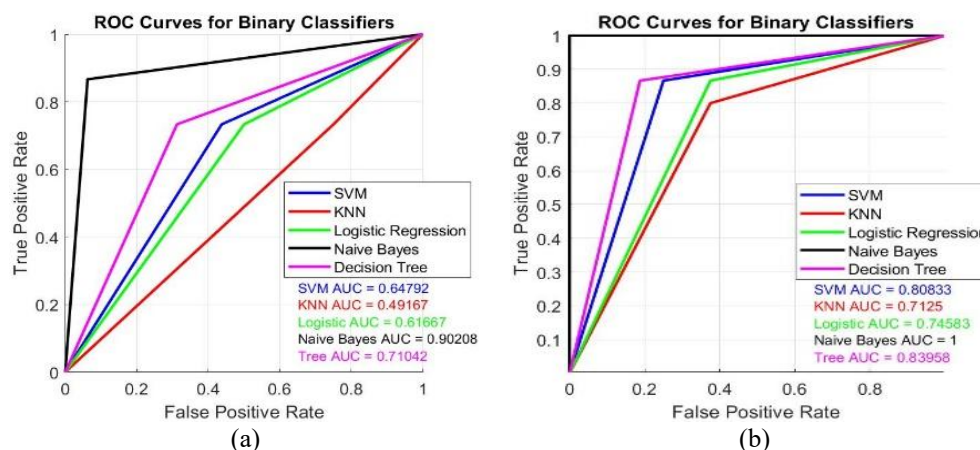


Figure 6. ROC curves for binary classifiers (a) conventional features and (b) proposed developed features

## 5. CONCLUSION

This study proposed a framework using in promoter detection by combining traditional nucleotide composition with newly developed features such as sequence complexity, variability, and k-mer-derived descriptors. Statistical analysis proved the importance of features such as Count\_C (AUC 0.7682), GC\_Content (AUC 0.7540), and Sequence\_Complexity (AUC 0.7380), which provided stable and discriminative signals for classification. The proposed feature set enhanced the overall performance of the model, resulting in an increased AUC when using the enhanced architecture. Among the five classifiers used in this study, the NB model obtained perfect results with an accuracy of 100%, a precision of 1.00, a recall of 1.00, and an F1-score of 1.00 when using enhanced features. These results confirmed that engineered features, based on biological and statistical properties of DNA sequences, can significantly enhance the classification performance even when simple models are used.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Sinan Salim Mohammed Sheet	✓	✓		✓			✓	✓	✓		✓		✓	
Marwa Mawfaq Mohamedsheet Al-Hatab	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓		✓	
Maysaloon Abed Qasim	✓	✓		✓	✓	✓	✓		✓	✓	✓	✓	✓	

- C : Conceptualization
- M : Methodology
- So : Software
- Va : Validation
- Fo : Formal analysis
- I : Investigation
- R : Resources
- D : Data Curation
- O : Writing - Original Draft
- E : Writing - Review & Editing
- Vi : Visualization
- Su : Supervision
- P : Project administration
- Fu : Funding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author, [SSMS], upon reasonable request.

REFERENCES

[1] S. R. Archuleta, J. A. Goodrich, and J. F. Kugel, "Mechanisms and functions of the RNA polymerase II general transcription machinery during the transcription cycle," *Biomolecules*, vol. 14, no. 2, Feb. 2024, doi: 10.3390/biom14020176.

[2] J. Yuan *et al.*, "A compendium of genetic variations associated with promoter usage across 49 human tissues," *Nature Communications*, vol. 15, no. 1, Oct. 2024, doi: 10.1038/s41467-024-53131-6.

[3] J. Blazeck and H. S. Alper, "Promoter engineering: recent advances in controlling transcription at the most fundamental level," *Biotechnology Journal*, vol. 8, no. 1, pp. 46–58, 2013, doi: 10.1002/biot.201200120.

[4] G. Brixi *et al.*, "Genome modeling and design across all domains of life with Evo 2," *bioRxiv*, 2025, doi: 10.1101/2025.02.18.638918.

[5] R. K. Umarov and V. V. Solovyev, "Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks," *PLoS one*, vol. 12, no. 2, Feb. 2017, doi: 10.1371/journal.pone.0171410.

[6] M. Oubounyt, Z. Louadi, H. Tayara, and K. T. Chong, "Deepromoter: robust promoter predictor using deep learning," *Frontiers in Genetics*, vol. 10, Apr. 2019, doi: 10.3389/fgene.2019.00286.

[7] Z.-W. Ma, J.-P. Zhao, J. Tian, and C.-H. Zheng, "DeeProPre: a promoter predictor based on deep learning," *Computational Biology and Chemistry*, vol. 101, 2022, doi: 10.1016/j.compbiolchem.2022.107770.

[8] W. Zhang *et al.*, "MethylGrapher: genome-graph-based processing of DNA methylation data from whole genome bisulfite sequencing," *Nucleic Acids Research*, vol. 53, no. 3, Jan. 2025, doi: 10.1093/nar/gkaf028.

[9] R. Amin *et al.*, "iPromoter-BnCNN: a novel branched CNN-based predictor for identifying and classifying sigma promoters," *Bioinformatics*, vol. 36, no. 19, pp. 4869–4875, Dec. 2020, doi: 10.1093/bioinformatics/btaa609.

[10] H. Tayara, M. Tahir, and K. T. Chong, "Identification of prokaryotic promoters and their strength by integrating heterogeneous features," *Genomics*, vol. 112, no. 2, pp. 1396–1403, Mar. 2020, doi: 10.1016/j.ygeno.2019.08.009.

[11] L. Moraes, P. Silva, E. Luz, and G. Moreira, "CapsProm: a capsule network for promoter prediction," *Computers in Biology and Medicine*, vol. 147, Aug. 2022, doi: 10.1016/j.compbiomed.2022.105627.

[12] Z. Zhang, J. Zhao, P.-J. Wei, and C.-H. Zheng, "iPromoter-CLA: Identifying promoters and their strength by deep capsule networks with bidirectional long short-term memory," *Computer Methods and Programs in Biomedicine*, vol. 226, 2022, doi: 10.1016/j.cmpb.2022.107087.

[13] Z. Li, J. Jin, W. Long, and L. Wei, "PLPMpro: enhancing promoter sequence prediction with prompt-learning based pre-trained language model," *Computers in Biology and Medicine*, vol. 164, Sep. 2023, doi: 10.1016/j.compbiomed.2023.107260.

[14] S. Paul, K. Olymon, G. S. Martinez, S. Sarkar, V. R. Yella, and A. Kumar, "MLDSPP: bacterial promoter prediction tool using DNA structural properties with machine learning and explainable AI," *Journal of Chemical Information and Modeling*, vol. 64, no. 7, pp. 2705–2719, Apr. 2024, doi: 10.1021/acs.jcim.3c02017.

[15] H. Ashayeri, N. Sobhi, P. Pławiak, S. Pedrammehr, R. Alizadehsani, and A. Jafarizadeh, "Transfer learning in cancer genetics, mutation detection, gene expression analysis, and syndrome recognition," *Cancers*, vol. 16, no. 11, Jun. 2024, doi: 10.3390/cancers16112138.




[16] R. Zeng, Z. Li, J. Li, and Q. Zhang, "DNA promoter task-oriented dictionary mining and prediction model based on natural language technology," *Scientific Reports*, vol. 15, no. 1, Jan. 2025, doi: 10.1038/s41598-024-84105-9.

[17] S. Gunarathna *et al.*, "ATAC-seq guided interpretable machine learning reveals cancer-specific chromatin features in cell-free DNA," *Research Square*, pp. 1–29, Jan. 2025, doi: 10.21203/rs.3.rs-5485170/v1.




- [18] Z. Lin, "Development of methods for cancer genome analysis and clinical applications," *Ph.D. thesis*, School of Arts and Sciences, Harvard University, Cambridge, Massachusetts, 2022.
- [19] S. Q. Hasan, "Shallow model and deep learning model for features extraction of images," *NTU Journal of Engineering and Technology*, vol. 2, no. 3, Nov. 2023, doi: 10.56286/ntujet.v2i3.449.
- [20] T. Jabid, F. Anwar, S. M. Baker, and M. Shoyaib, "Identification of promoter through stochastic approach," in *2007 10th International Conference on Computer and Information Technology, ICCIT*, Dec. 2007, pp. 1–4, doi: 10.1109/ICCITECHN.2007.4579366.
- [21] R. H. M. Ameen, N. M. Basheer, and A. K. Younis, "A survey: breast cancer classification by using machine learning techniques," *NTU Journal of Engineering and Technology*, vol. 2, no. 1, May 2023, doi: 10.56286/ntujet.v2i1.367.
- [22] I. Sarker, M. Faruque, H. Alqahtani, and A. Kalim, "K-nearest neighbor learning based diabetes mellitus prediction and analysis for eHealth services," *EAI Endorsed Transactions on Scalable Information Systems*, vol. 7, no. 6, pp. 1–9, Jul. 2018, doi: 10.4108/eai.13-7-2018.162737.
- [23] R. D. Joshi and C. K. Dhakal, "Predicting type 2 diabetes using logistic regression and machine learning approaches," *International Journal of Environmental Research and Public Health*, vol. 18, no. 14, 2021, doi: 10.3390/ijerph18147346.
- [24] M. S. Alzboon, M. Alqaraleh, and M. S. Al-Batah, "Diabetes prediction and management using machine learning approaches," *Data and Metadata*, vol. 4, 2025, doi: 10.56294/dm2025545.
- [25] K. Jhaharia and P. Mathur, "A comprehensive review on machine learning in agriculture domain," *IAES International Journal of Artificial Intelligence*, vol. 11, no. 2, pp. 753–763, Jun. 2022, doi: 10.11591/ijai.v11.i2.pp753-763.
- [26] Y. J. Lee, S. W. Kweon, C. W. Jeong, and H. J. Kim, "Evaluating the performance of machine learning and variable selection methods to identify document paper using infrared spectral data," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 327, Feb. 2025, doi: 10.1016/j.saa.2024.125299.
- [27] O. Rainio, J. Teuhio, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Scientific Reports*, vol. 14, no. 1, Mar. 2024, doi: 10.1038/s41598-024-56706-x.
- [28] R. Diallo, C. Edalo, and O. O. Awe, "Machine learning evaluation of imbalanced health data: a comparative analysis of balanced accuracy, MCC, and F1 score," in *Practical Statistical Learning and Data Science Methods: Case Studies from LISA 2020 Global Network, USA*, Springer, 2025, pp. 283–312, doi: 10.1007/978-3-031-72215-8\_12.

## BIOGRAPHIES OF AUTHORS






**Sinan Salim Mohammed Sheet**    holds a master's and doctor of Biomedical Engineering from University Teknologi Malaysia (UTM) University, Malaysia in 2011 and 2022 respectively. He also received his B.Sc. in Medical Instrumentation Engineering from the Technical College of Mosul, Iraq in 1999. He is currently an Associate Professor at the Department of Instrument Engineering in Mosul, Northern Technical University, Mosul, Iraq. He can be contacted at email: sinan\_sm76@ntu.edu.iq or sinan\_sm76@yahoo.com.



**Marwa Mawfaq Mohamedsheet Al-Hatab**    received the B.Sc. degree in Medical Instruments Engineering from Technical Engineering College, Mosul, Northern Technical University (NTU), Mosul, Iraq, and the M.Sc. degree in Biomedical Engineering from Università Politecnica delle Marche (UNIVPM). She is a lecturer at Technical Engineering college, Mosul, Northern Technical University (NTU), Mosul, Iraq. She can be contacted at email: marwa.alhatab@ntu.edu.iq.



**Maysaloon Abed Qasim**    received the B.Sc. degree in Electrical Engineering from Mosul University, Iraq, the M.Sc. degree in Electronic and Communication from Mosul University, Iraq, and the Ph.D. degree Electronic and Computer Engineering from Altinbas University, Istanbul, Turkey. She is Assistant Professor at Technical Engineering College, Northern Technical University (NTU), Mosul, Iraq. She can be contacted at email: maysaloon.alhashim@ntu.edu.iq.