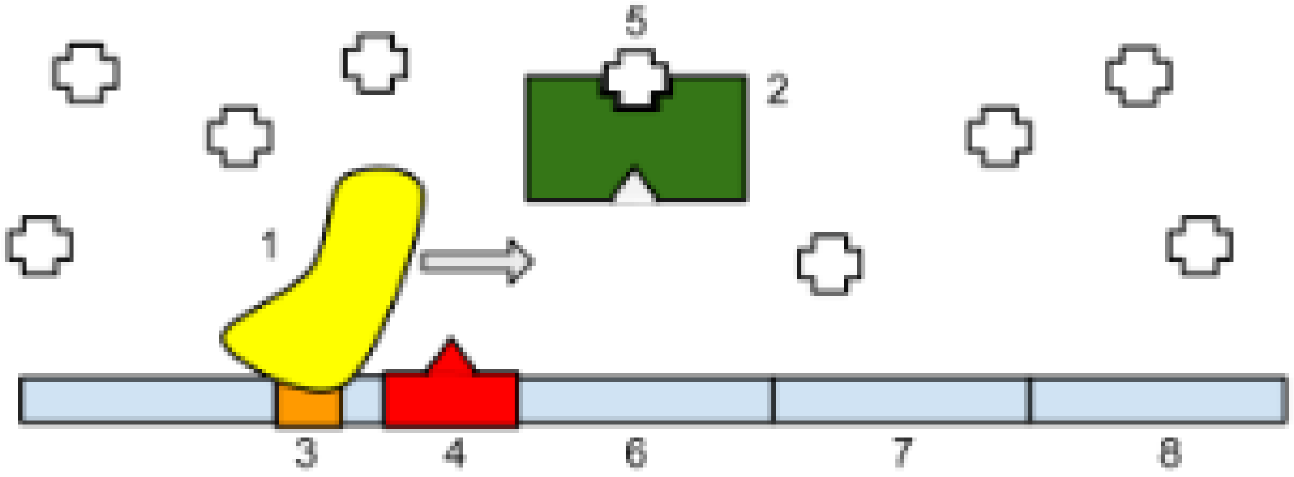
Sigma Promoter Identification & Classification in E. coli bacteria from Genomic Data using Deep Learning based Approach

Presented by,

Rifat Rahman (0419052028)

Promoter

- **Promoter** is a sequence of DNA to which proteins bind that initiate transcription of a single RNA from the DNA downstream of it.
- For transcription occurrence, RNA polymerase must bind near the promoter.
- Promoters can be about 100–1000 base pairs long.



1: **RNA Polymerase,** ***2***: **Repressor,** ***3***: **Promoter,** ***4***: **Operator,** ***5***: **Lactose,** ***6***: **lacZ,** ***7***: **lacY,** ***8***: **lacA.**

Fig: Transcription occurring from DNA to RNA

Sigma Factor

There are several sigma factors in the RNA polymerase of *Escherichia coli* bacteria, which are dependent on environment & gene. These sigma factors are used as distinguishing elements of promoter sequences found in DNA.

Sigma Factors	Functions
σ70	Responsible for transcription of most of the genes under normal condition
σ24	Responsible for heat shock response
σ28	Responsible for flagellar genes
σ32	Responsible for heat shock response
σ38	Responsible for stress response during the transition from exponential growth phase to the stationary phase of E. coli
σ54	Responsible for nitrogen metabolism

Table: Tasks of different sigma factor of *E. coli* bacteria

Problem Definition & Motivation

Problem Definition

Input: Nucleotide sequence and structural characteristics of DNA
output: Identification and classification of sigma factor dependent promoters

Motivation

- Promoter is responsible for initiating transcription of specific genes. So, it is essential to build automatic identifier and classifier to detect & classify promoters.
- Molecular techniques for promoter identification or classification is costly in terms of time and money.
- Promoters may have both intra and inter class variation and similarity in terms of consensus sequences and positions.
- Accurate classification of various types of sigma promoters still remains a challenge.

Related Works (iPromoter-FSEn)

Rahman et al. [1] developed iPromoter-FSEn for identifying bacterial σ70 promoter using feature subspace based ensemble classifier achieving an impressive accuracy of 86.32%.

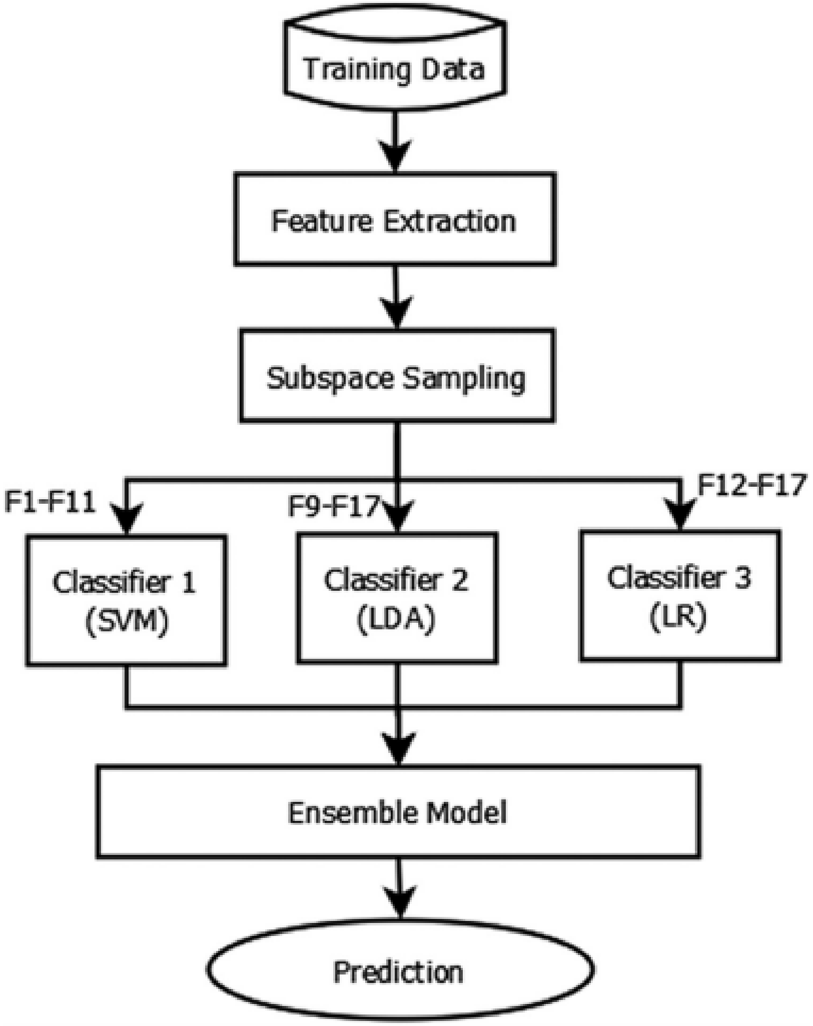


Fig: System diagram of iPromoter-FSEn

Related Works (iPromoter-2L)

Liu et al. [2] implemented iPromoter-2L where they extracted physiochemical properties by multi-window-based PseKNC. Authors used random forest classifier for classification.

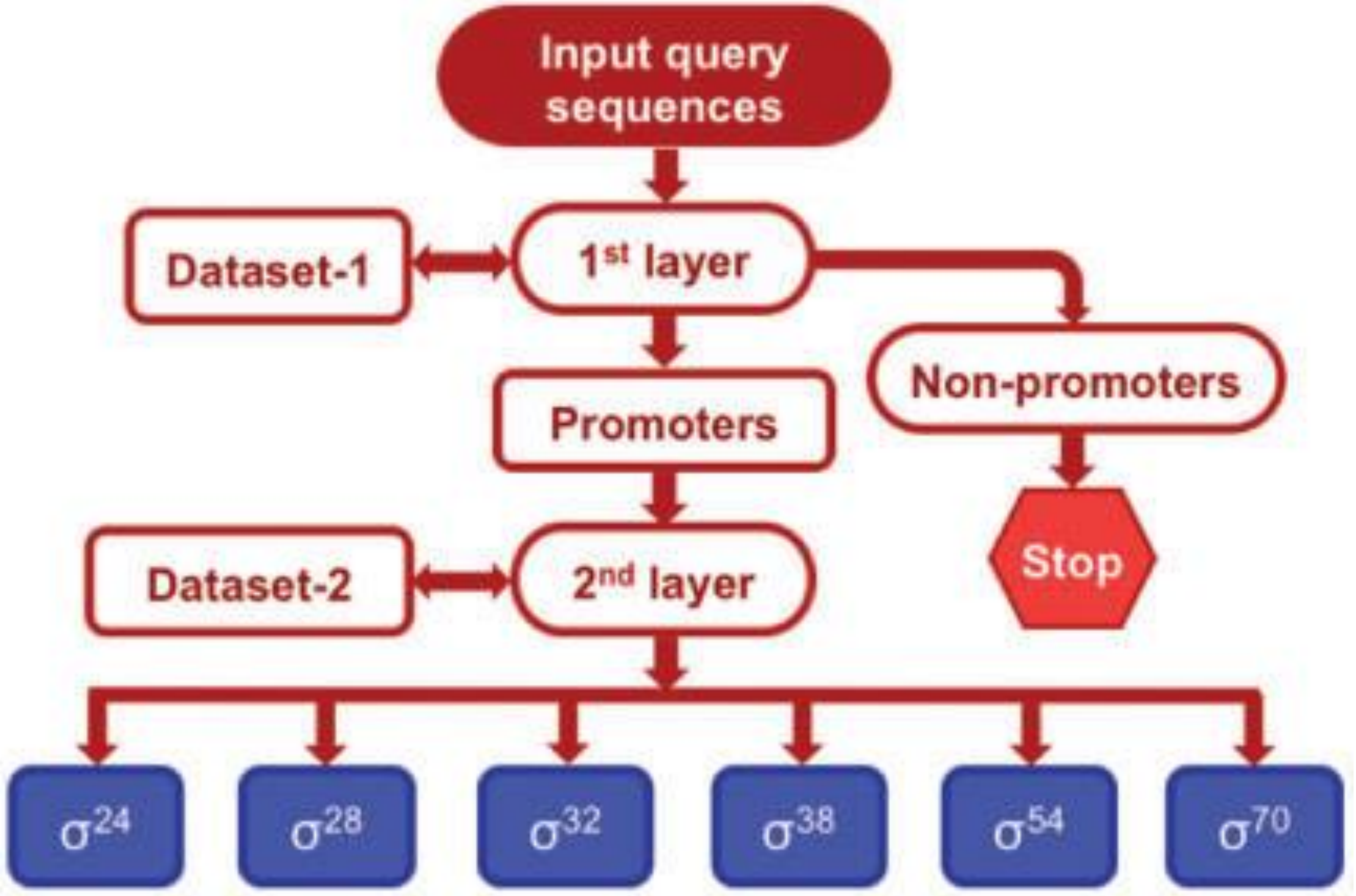


Fig: A flow chart to show how the iPromoter-2L is working

Related Works (MULTiPly)

Zhang et al. [3] extracted both local (k-tuple nucleotide composition, dinucleotide based auto covariance) and global information (bi-profile Bayes and KNN feature encodings), and performed F-score feature selection method to identify the best unique type of feature prediction results. They used SVM classifier for classification.

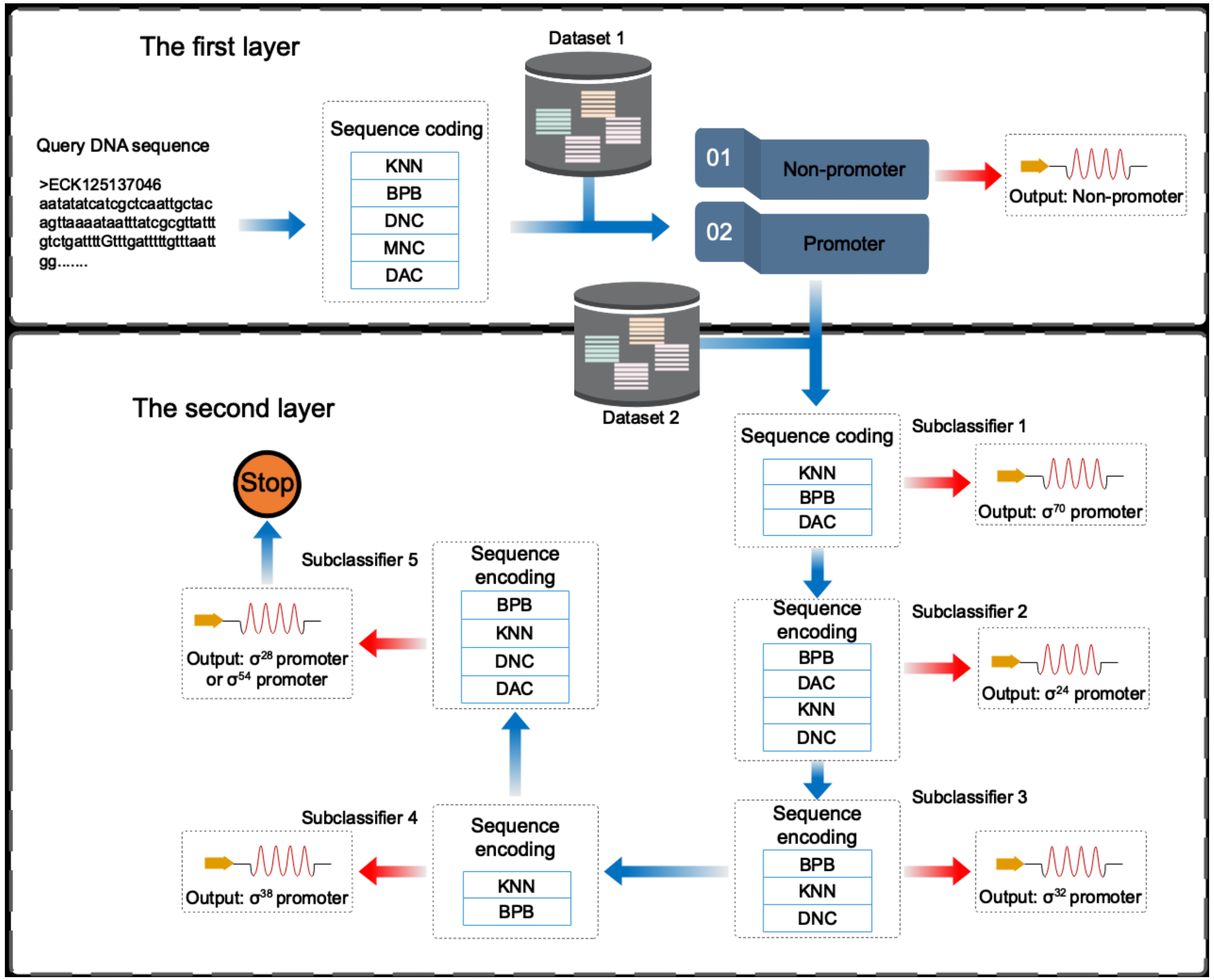


Fig: The flowchart of the proposed multi-layer classifier

Related Works (iPromoter-BnCNN)

Amin et al. [4] combines local features related to mono-mer nucleotide sequence, tri-mer nucleotide sequence, di-mer structural properties and tri-mer structural properties through the use of parallel branching.

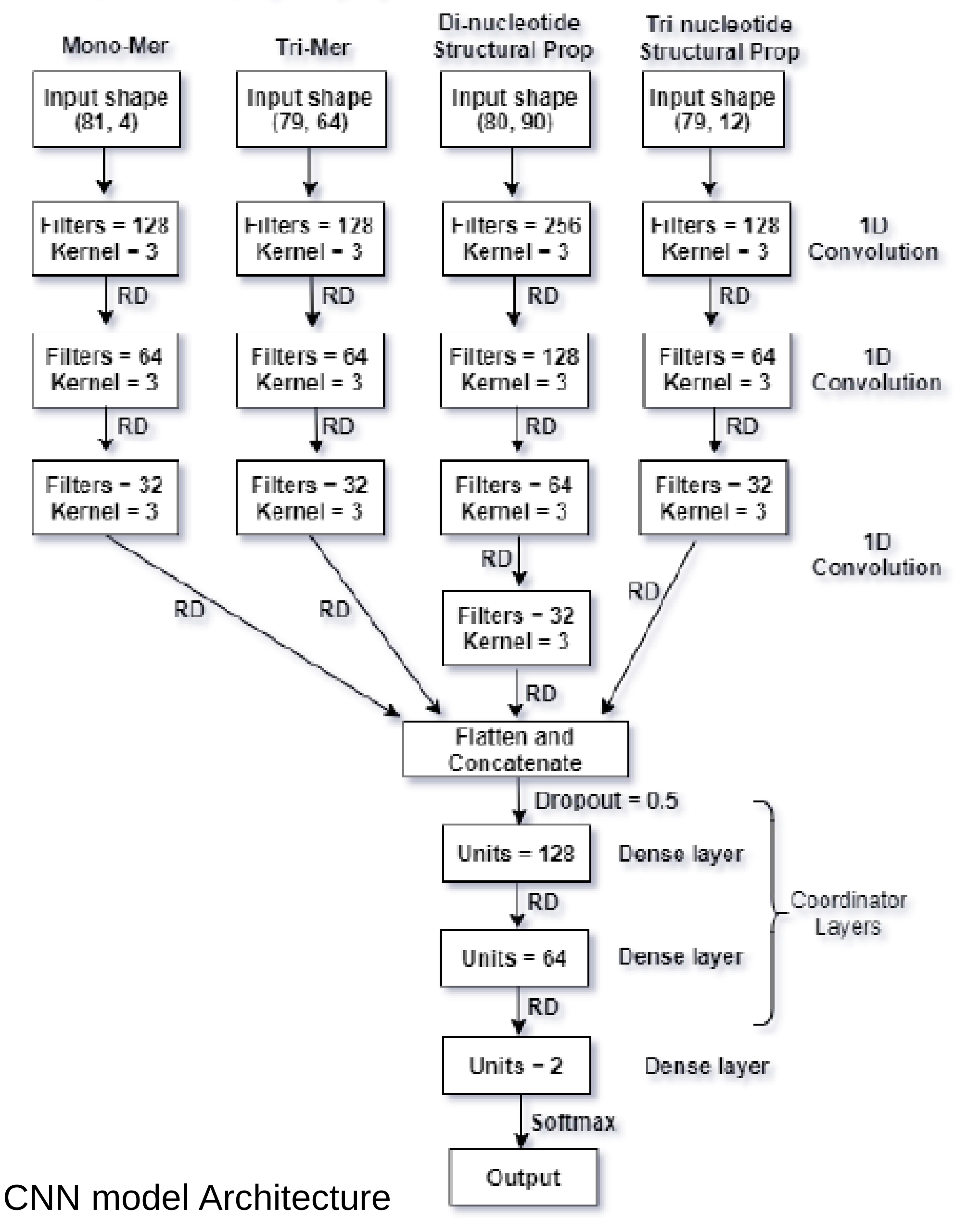


Fig: Parallel Branched CNN model Architecture

Related Works (iPromoter-BnCNN)

Potential Motif identification:

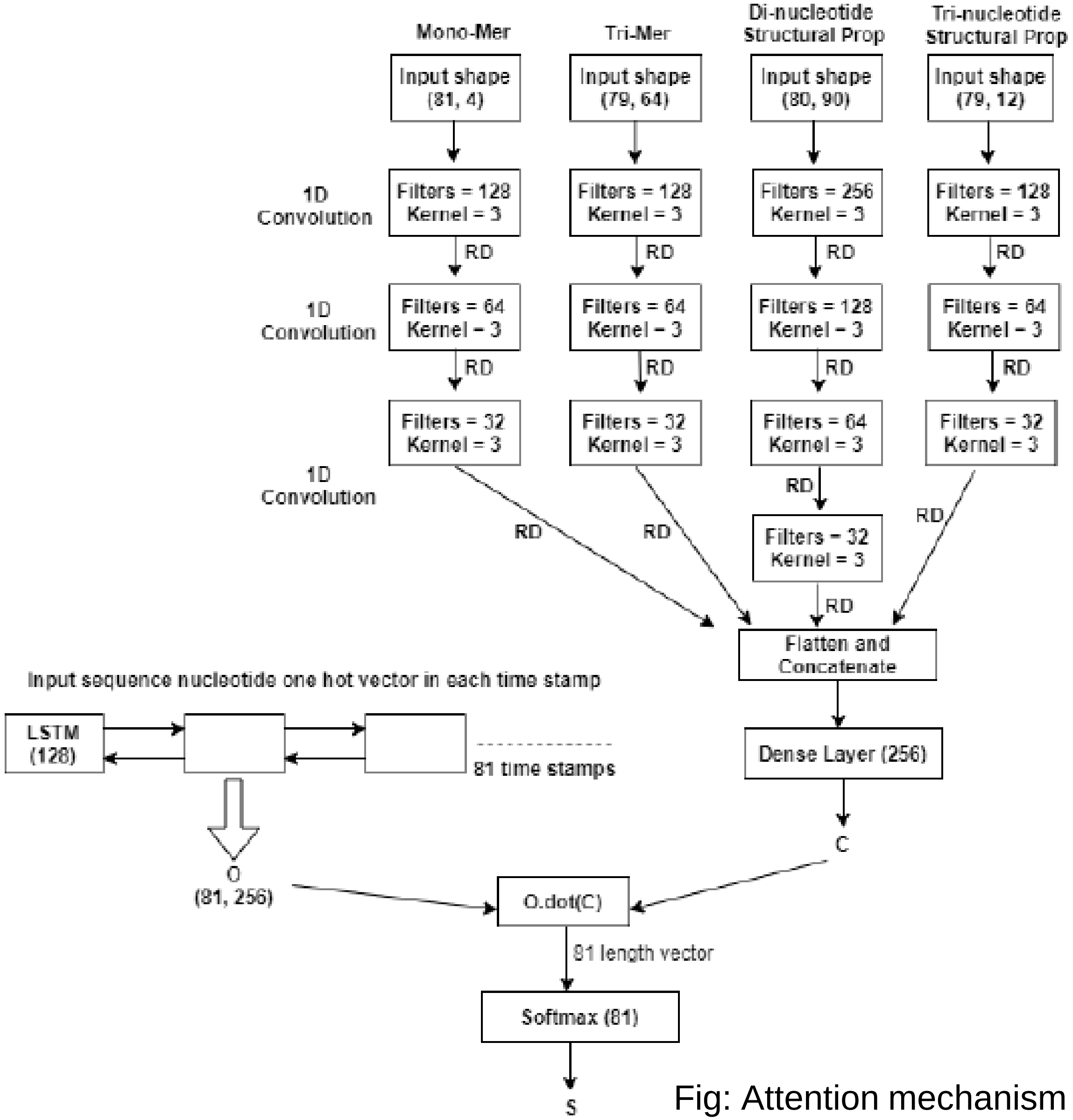
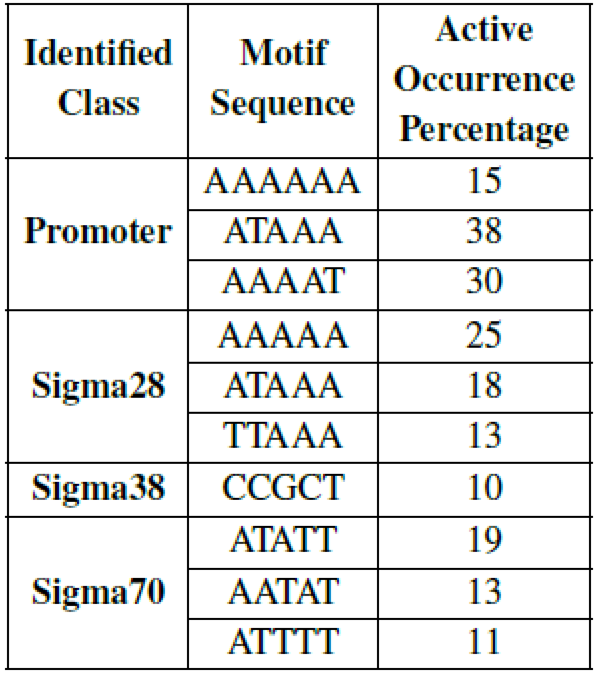


Fig: Attention mechanism for motif identification



Identified Class	Motif Sequence	Active Occurrence Percentage
Promoter	AAAAAA	15
	ATAAA	38
	AAAAT	30
Sigma28	AAAAA	25
	ATAAA	18
	TTAAA	13
Sigma38	CCGCT	10
Sigma70	ATATT	19
	AATAT	13
	ATTTT	11

Table: Identified potential motifs using attention mechanism

Related Works

- Silva et al. [5] integrated DNA duplex stability as feature of neural network to identify σ28 and σ54 class of promoter in E. coli bacteria.

- Umarov and Solovyev [6] trained CNN based architecture to Recognize prokaryotic and eukaryotic promoters.

Gap Analysis

- Most of the works implemented identifiers for one or two sigma factor dependent promoters except [2-4]

- The sensitivity and specificity of promoter classification showed opposing behavior for iPromoter-2L [2]. For example, for σ28, σ32, σ38, and σ54, iPromoter-2L showed specificity of higher than 99%, but the sensitivity was lower than 54%.

- The limitation of MULTiPly [3] was the selection of the basic features to work with. Different combination of different heterogeneous features led to different prediction results. Through trial and error, the authors selected features that achieved satisfactory prediction performance.

- Implementation of multiple binary classifiers.

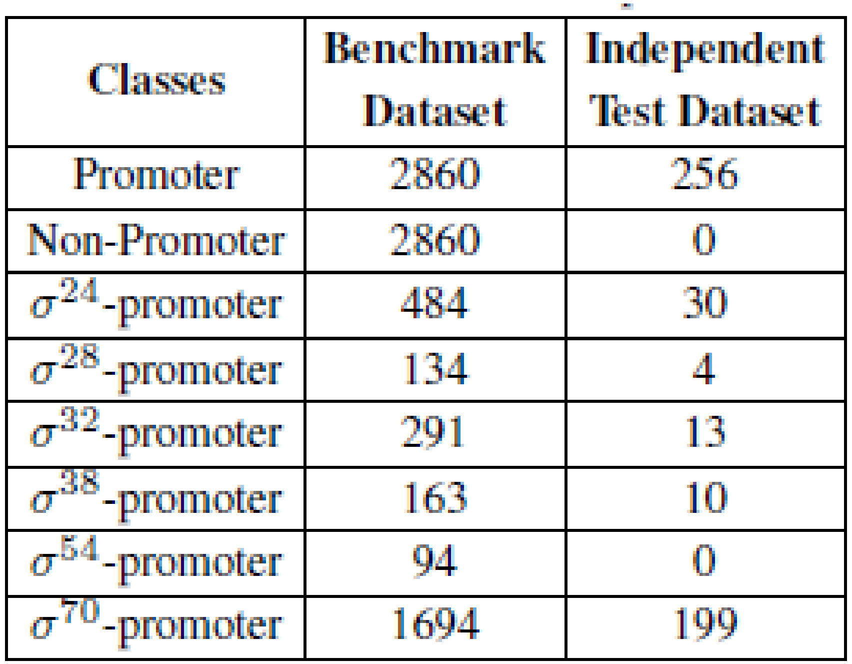
Proposed Method

Objective

Our goal is to implement a computational tool that will identify and classify promoters in E. coli bacteria automatically. We will also perform comparative study among different state-of-the-art works.

Benchmark Dataset Selection

- Redundancy reduced “RegulonDB version-9.3”
- Creating training and independent test dataset
- DNA sequence length: 81 nucleotides



Classes	Benchmark Dataset	Independent Test Dataset
Promoter	2860	256
Non-Promoter	2860	0
σ^{24}-promoter	484	30
σ^{28}-promoter	134	4
σ^{32}-promoter	291	13
σ^{38}-promoter	163	10
σ^{54}-promoter	94	0
σ^{70}-promoter	1694	199

Table: Class-wise sample numbers in datasets used

Proposed Method

Mathematical Formulation of DNA Sequence

- Original Nucleotide sequence
  - Mono-mer (81 x 4D)
  - Di-mer (80 x 16D)
  - Tri-mer (79 x 64D)

- Structural Properties
  - Di-mer (80 x 90D)
  - Tri-mer (79 x 12D)

Here, rows are representing polymers from input sequence and columns are representing features.

Dimensionality Reduction and clustering (Our Novelty)

- Auto-encoder/t-SNE/PCA
- Clustering for purity checking

Proposed Method

Model Architecture (Our Novelty)

- Sequence Model
- Bi-LSTM based attention mechanism

Evaluation

- Hyper-parameter Tuning
  - Number of hidden layers
  - Number of neurons
  - Function
  - Learning rate

- Performance Metrics
  - Sensitivity
  - Specificity
  - Accuracy
  - MCC score

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

$$Sn = \frac{TP}{TP + FN}$$

$$Sp = \frac{TN}{TN + FP}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Proposed Method

- **Tools**
  - Platform: Google colab
  - Language: Python
  - Library: keras with tensorflow background, scikit-learn, numpy, pandas, matplotlib etc

- **Public Access**
  - We will keep source codes & model files public.

Proposed Method (Block Diagram)

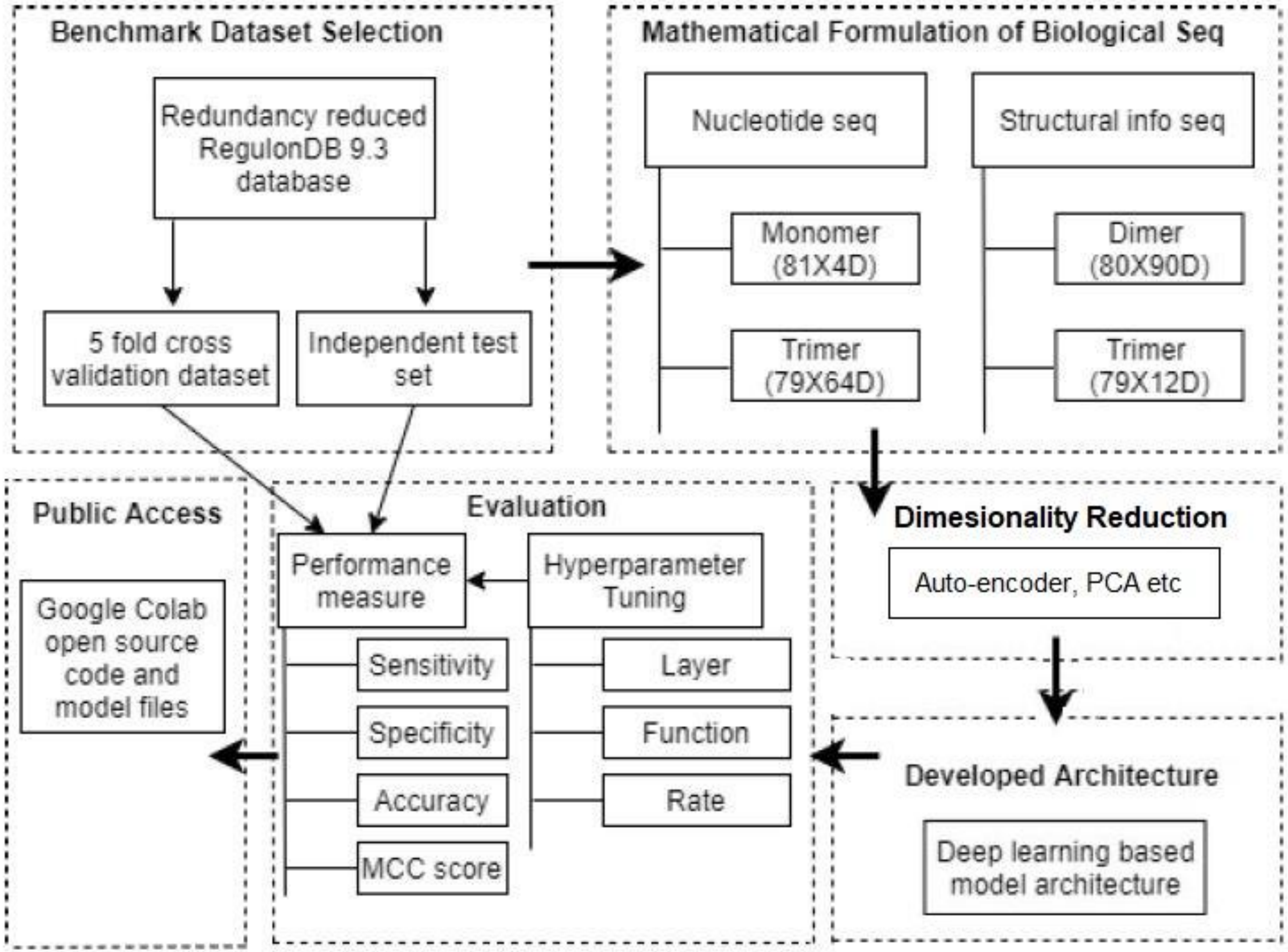


Fig: Block diagram of our methodology

References

[1] Rahman, M. S. et al. (2019b). ipromoter-fsen: Identification of bacterial σ70 promoter sequences using feature subspace based ensemble classifier. Genomics, 111(5), 1160–1166.

[2] M B. Liu, F. Yang, D.-S. Huang, and K.-C. Chou, "ipromoter-2l: a two-layer predictor for identifying promoters and their types by multi-window-based pseknc," Bioinformatics, vol. 34, no. 1, pp. 33-40, 2018.

[3] M. Zhang, F. Li, T. T. Marquez-Lago, A. Leier, C. Fan, C. K. Kwoh, K.-C. Chou, J. Song, and C. Jia, "Multiply: a novel multi-layer predictor for discovering general and specific types of promoters," Bioinformatics, vol. 35, no. 17, pp. 2957-2965, 2019.

[4] R. Amin, C. R. Rahman, S. Ahmed, M. Sifat, H. Rahman, M. N. K. Liton, M. Rahman, M. Khan, Z. Hossain, S. Shatabda et al., "ipromoter-bncnn: a novel branched cnn based predictor for identifying and classifying sigma promoters," Bioinformatics, 2019.

[5] S. d. A. e Silva, F. Forte, I. T. Sartor, T. Andrighetti, G. J. Gerhardt, A. P. L. Delamare, and S. Echeverrigaray, "DNA duplex stability as discriminative characteristic for *escherichia coli* σ54-and σ28-dependent promoter sequences," Biologicals, vol. 42, no. 1, pp. 22-28, 2014.

[6] 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, p. e0171410, 2017.

Thank You