piEnPred: a bi-layered discriminative model for enhancers and their subtypes via novel cascade multi-level subset feature selection algorithm

Zaheer Ullah KHAN, Dechang PI , Shuanglong YAO, AsifNAWAZ, Farman ALI, Shaukat ALI

Frontiers of Computer Science, DOI: 10.1007/s11704-020-9504-3

Problems Formulation & Proposed Model

Problem statement

Enhancers are short DNA cis-elements that can be bound by proteins (activators) to increase the possibility that transcription of a particular gene will occur.

The Enhancers perform a significant role in the

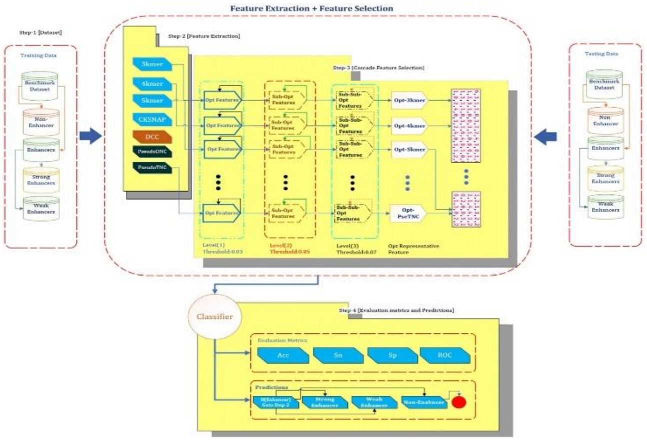
- Formation of proteins
- Regulating the gene transcription process.
- Human diseases such as cancer, inflammatory bowel disease, Parkinson’s, addiction, and schizophrenia are due to genetic variation in enhancers

Proposed Model

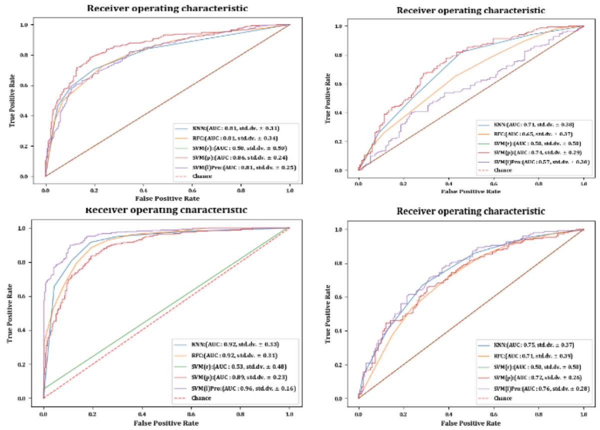
- A novel computational a bi-layered model is developed for the prediction of Enhancers and their subtypes. The feature vector was constructed over a linear combination of six features. The optimum Hybrid feature vector was obtained via the Novel Cascade Multi-Level Subset Feature selection (CM-SFS) algorithm

Main Contributions

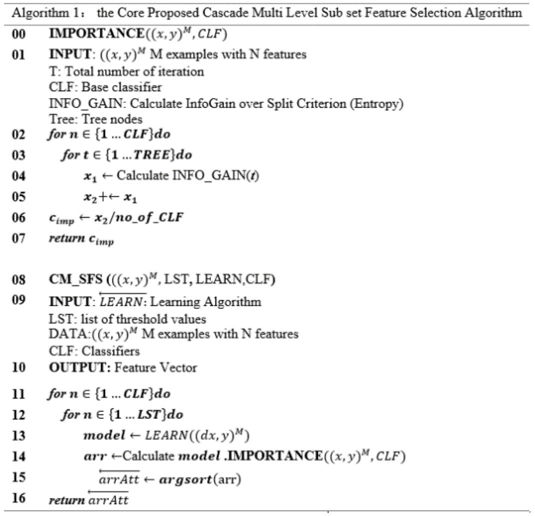
- **Proposed Model framework**



- **piEnPred improvements over 1st & 2nd Layers**



- **Algorithm Pseudo code**



Algorithm 1: the Core Proposed Cascade Multi Level Sub set Feature Selection Algorithm

```
IMPORTANCE($(x,y)^M, CLF$)
INPUT: $((x,y)^M$ M examples with N features
T: Total number of iteration
CLF: Base classifier
INFO_GAIN: Calculate InfoGain over Split Criterion (Entropy)
Tree: Tree nodes
for $n \in \{1 \dots CLF\}$ do
  for $t \in \{1 \dots TREE\}$ do
      $x_1 \leftarrow$ Calculate INFO_GAIN($t$)
      $x_2 + \leftarrow x_1$
$c_{imp} \leftarrow x_2 / no\_of\_CLF$
return $c_{imp}$

CM_SFS ($((x,y)^M$, LST, LEARN,CLF)
INPUT: $\overleftarrow{LEARN}$: Learning Algorithm
LST: list of threshold values
DATA:$((x,y)^M$ M examples with N features
CLF: Classifiers
OUTPUT: Feature Vector
for $n \in \{1 \dots CLF\}$ do
  for $n \in \{1 \dots LST\}$ do
      $model \leftarrow LEARN((dx,y)^M)$
      $arr \leftarrow$Calculate $model$ .IMPORTANCE($(x,y)^M, CLF$)
        $\overleftarrow{arrAtt} \leftarrow argsort$(arr)
return $\overleftarrow{arrAtt}$
```

- **Model Comparison**

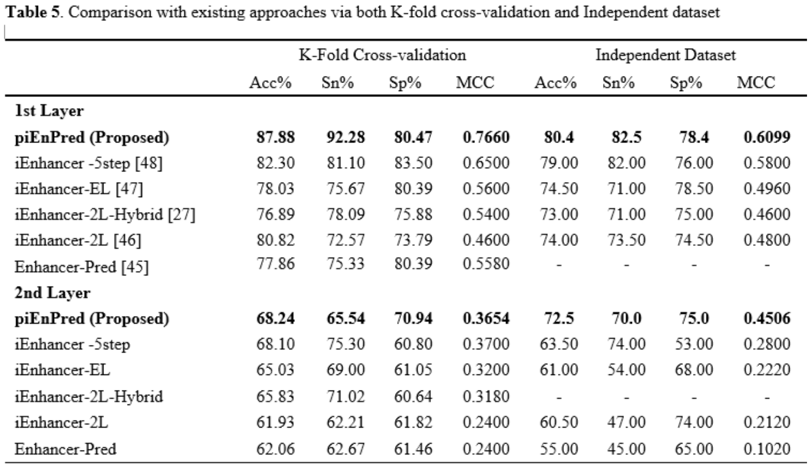


Table 5. Comparison with existing approaches via both K-fold cross-validation and Independent dataset

	K-Fold Cross-validation				Independent Dataset			
	Acc%	Sn%	Sp%	MCC	Acc%	Sn%	Sp%	MCC
1st Layer								
piEnPred (Proposed)	**87.88**	**92.28**	**80.47**	**0.7660**	**80.4**	**82.5**	**78.4**	**0.6099**
iEnhancer -5step [48]	82.30	81.10	83.50	0.6500	79.00	82.00	76.00	0.5800
iEnhancer-EL [47]	78.03	75.67	80.39	0.5600	74.50	71.00	78.50	0.4960
iEnhancer-2L-Hybrid [27]	76.89	78.09	75.88	0.5400	73.00	71.00	75.00	0.4600
iEnhancer-2L [46]	80.82	72.57	73.79	0.4600	74.00	73.50	74.50	0.4800
Enhancer-Pred [45]	77.86	75.33	80.39	0.5580	-	-	-	-
2nd Layer								
piEnPred (Proposed)	**68.24**	**65.54**	**70.94**	**0.3654**	**72.5**	**70.0**	**75.0**	**0.4506**
iEnhancer -5step	68.10	75.30	60.80	0.3700	63.50	74.00	53.00	0.2800
iEnhancer-EL	65.03	69.00	61.05	0.3200	61.00	54.00	68.00	0.2220
iEnhancer-2L-Hybrid	65.83	71.02	60.64	0.3180	-	-	-	-
iEnhancer-2L	61.93	62.21	61.82	0.2400	60.50	47.00	74.00	0.2120
Enhancer-Pred	62.06	62.67	61.46	0.2400	55.00	45.00	65.00	0.1020