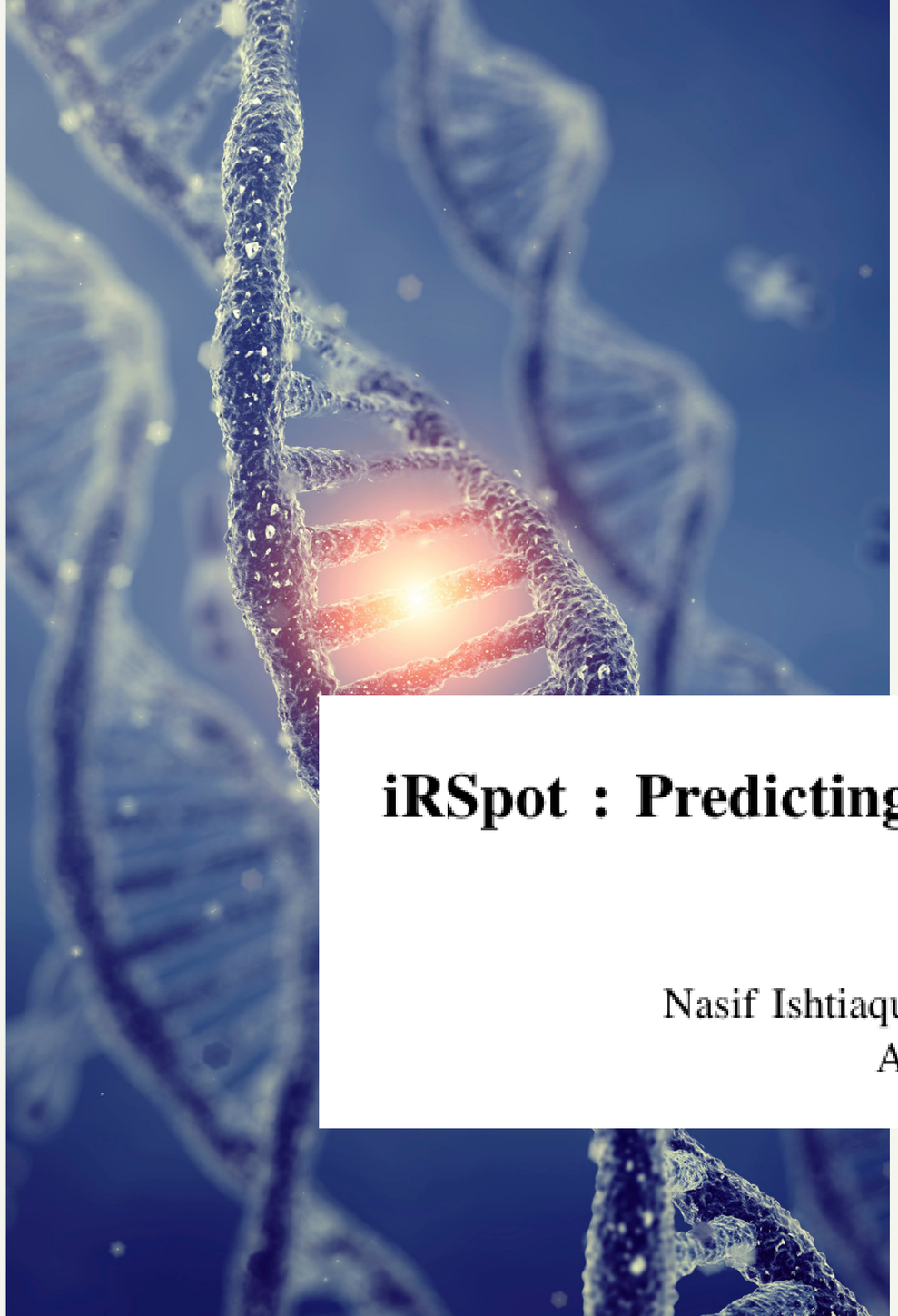


CSI-416
Pattern Recognition Laboratory
Final Presentation

Team name: black_crow

iRSpot : Predicting sequence based recombination hotspot using CNN-1D

Nasif Ishtiaque Islam [1] and Shayed Ashraf[1] and Ashak Mahmood[1]
Anika Tabassum[1] and Md. Rakibul Haque[2]

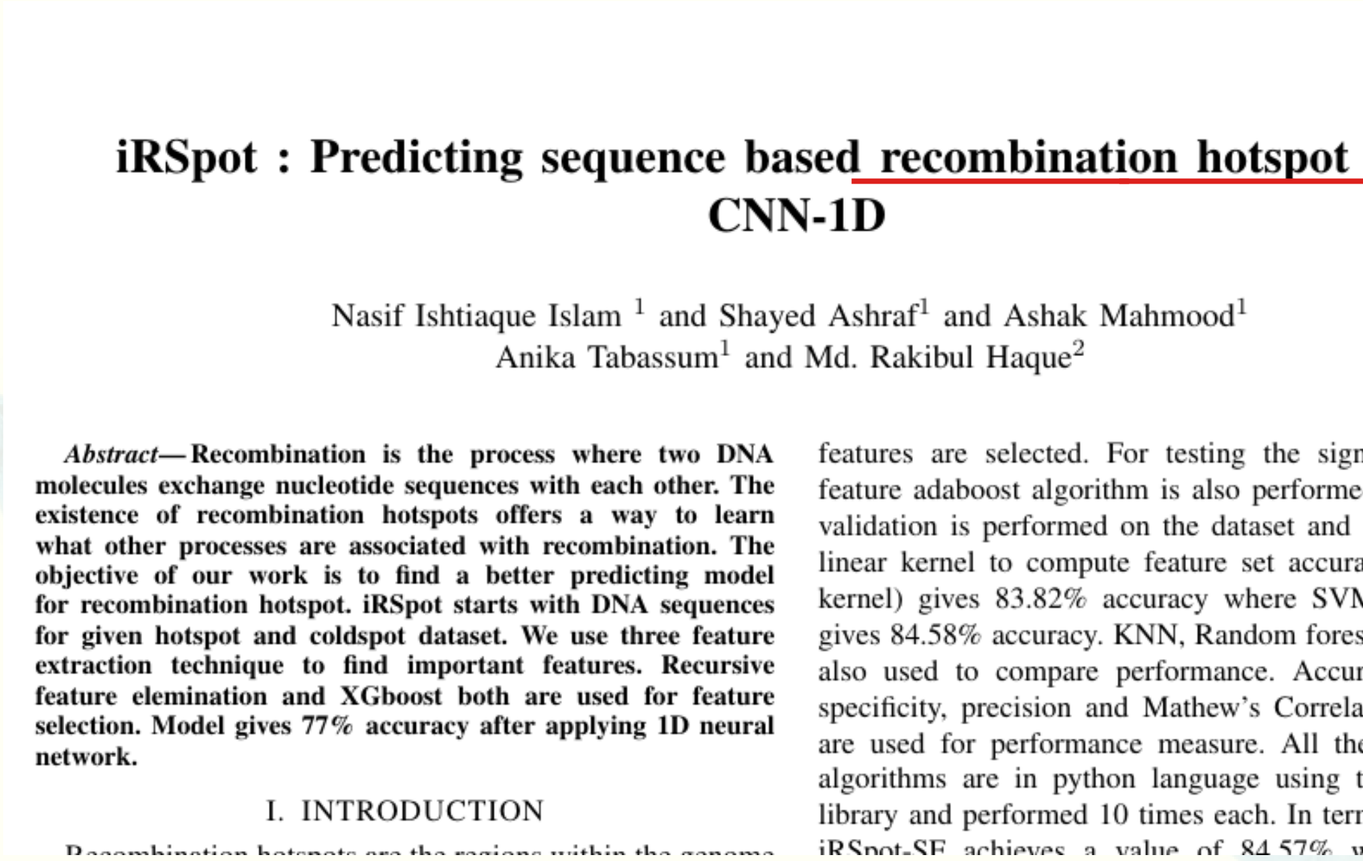
Introduction

What is recombination?

- Recombination is the process where two DNA molecules exchange nucleotide sequences with each other.

Importance of recombination:

- Recombination provides knowledge about DNA sequence variation and patterns along human chromosomes and this may help to map the position of alleles that cause various diseases.
- Recombination hotspot gives useful insights into the basic function of inheritance and the study of genetic diversity.



iRSpot : Predicting sequence based recombination hotspot CNN-1D

Nasif Ishtiaque Islam [1] and Shayed Ashraf[1] and Ashak Mahmood[1]
Anika Tabassum[1] and Md. Rakibul Haque[2]

***Abstract*— Recombination is the process where two DNA molecules exchange nucleotide sequences with each other. The existence of recombination hotspots offers a way to learn what other processes are associated with recombination. The objective of our work is to find a better predicting model for recombination hotspot. iRSpot starts with DNA sequences for given hotspot and coldspot dataset. We use three feature extraction technique to find important features. Recursive feature elemination and XGboost both are used for feature selection. Model gives 77% accuracy after applying 1D neural network.**

I. INTRODUCTION

features are selected. For testing the sign
feature adaboost algorithm is also performe
validation is performed on the dataset and
linear kernel to compute feature set accura
kernel) gives 83.82% accuracy where SVN
gives 84.58% accuracy. KNN, Random fores
also used to compare performance. Accur
specificity, precision and Mathew's Correla
are used for performance measure. All the
algorithms are in python language using t
library and performed 10 times each. In tern

Figure 1: Report Front Page

Introduction

Hotspot and Coldspot?

In genomic regions:

Hotspots ⟶ higher frequencies of recombination
Coldspots ⟶ lower frequencies of recombination

Recombination hotspot plays a vital part in evolutionary development.

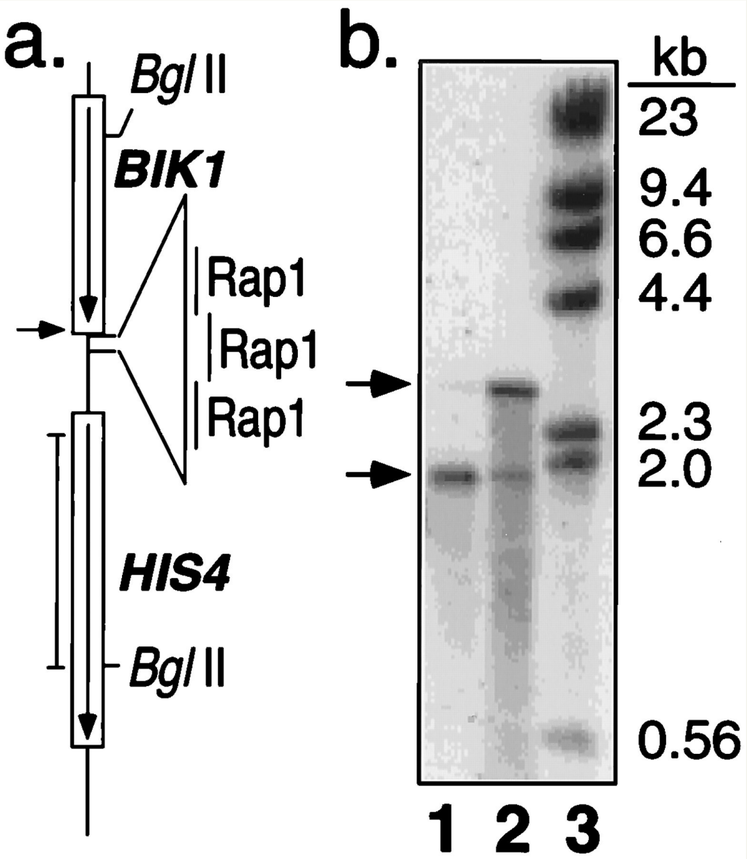


Figure 2: Hotspot-Coldspot *(Image from Internet)*

Methodology

- Three methods are used for feature extraction on our project: **K-mer composition, Gapped K-mer composition, Reverse complement K-mer composition.**
- Used **XGBoost and Recursive Feature Elimination** (Decision Tree Classifier) technique to eliminate less important features.
- **SVC, Gaussian NB, RF, AdaBoost, LR, KNN, Decision Tree and CNN(1D)** algorithms are also used to compare performance.
- All the programs and algorithms are in python language using the sci-kit learn library. In terms of **specificity our model achieves a value of 89.97%.**

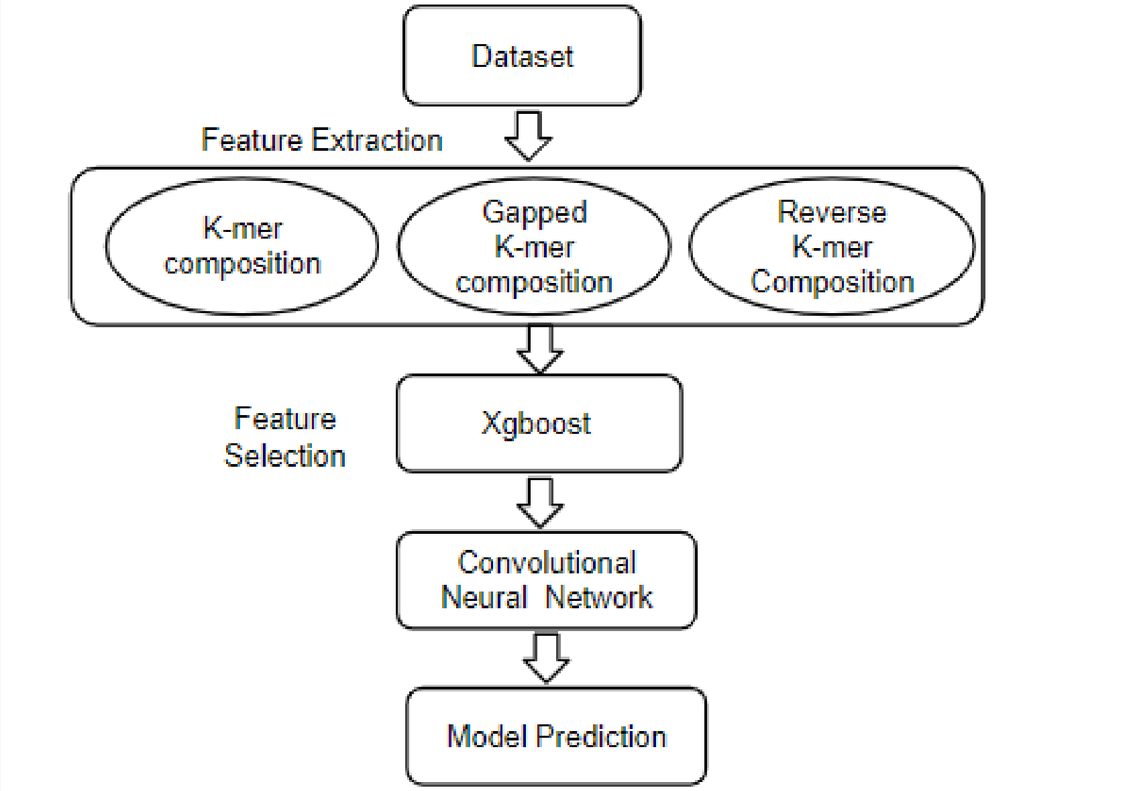


Figure 3: Model Architecture

Methodology - (Feature Extraction)

Dataset Description: The datasets used here is yeast dataset consisting of DNA sequences of nucleotides with both positive and negative instances. The positive instances are denoted as hotspot and negative are as coldspots. Dataset has 490 DNA segments of hotspot samples(positive) and 591 DNA segments(negative) of coldspot samples. The basic symbols of DNA sequences are A, T, C, G. This dataset represents the set of these sequences. Dataset is slightly imbalanced with less number of positive samples.

1. **K-mer:** K-mer is the substring of any length k in a sequence. Counting K-mer is an essential technique in many bioinformatics methods. In our p, we calculated till 4-mer and extracted 340 features from K-mer. Then we calculated the occurrence of specific K-mer.
2. **Gapped K-mer :** In order to find a tradeoff between the sparse feature space problem and more sequence composition information, the gapped K-mer has been proposed. Gapped K-mer allows several gaps to exist in K-mers. We extracted till 5 gaps, total 80 features using this technique.
3. **Reverse K-mer:** Counting K-mer on opposite direction is called Reverse K-mer. In our model we used till 4-mer in reverse order and extracted 340 feature. It's also detect the genome sequences of hidden patterns.

- Using these three techniques we extracted total 760 features from our datasets.

Methodology - (Feature Selection)

Figure 4: Feature Selection *(Image from Internet)*

- **XGBoost:** XGBoost is an implementation of gradient boosted decision trees designed for speed and performance. It is a feature selection technique to reduce unnecessary features. In our implementation we used Xgboost and selected top 40% (304) features and XGBoost classifier gives 74.46% accuracy on selected features.

- **Recursive Feature elimination (RFE):** Recursive feature elimination (RFE) is a feature selection method that removes the insignificant features according to their importance and then remove them recursively. Among 760 features we selected 50% (380) features using RFE and Desicion Tree Classifier gives 69% accuracy on selected features.

Methodology - (Algorithms)

After feature selection, we performed rigorous experiments on the dataset for selecting the most robust classification algorithm for our feature set. **Eight different classifiers** were investigated in our experi-ments: Decision Tree (DT) , Logistic Regression (LR), K-Nearest Neighbor (KNN), AdaBoost, Gaussian Naive Bayes (NB), Random Forest (RF), Support Vector Classifier with rbf kernel(SVC) and 1D-Convolutional Neural Network(CNN-1D).

In terms of Accuracy, worst performed classifier was Decision Tree and **best performed classifier is CNN-1D**. (*figure-5)*

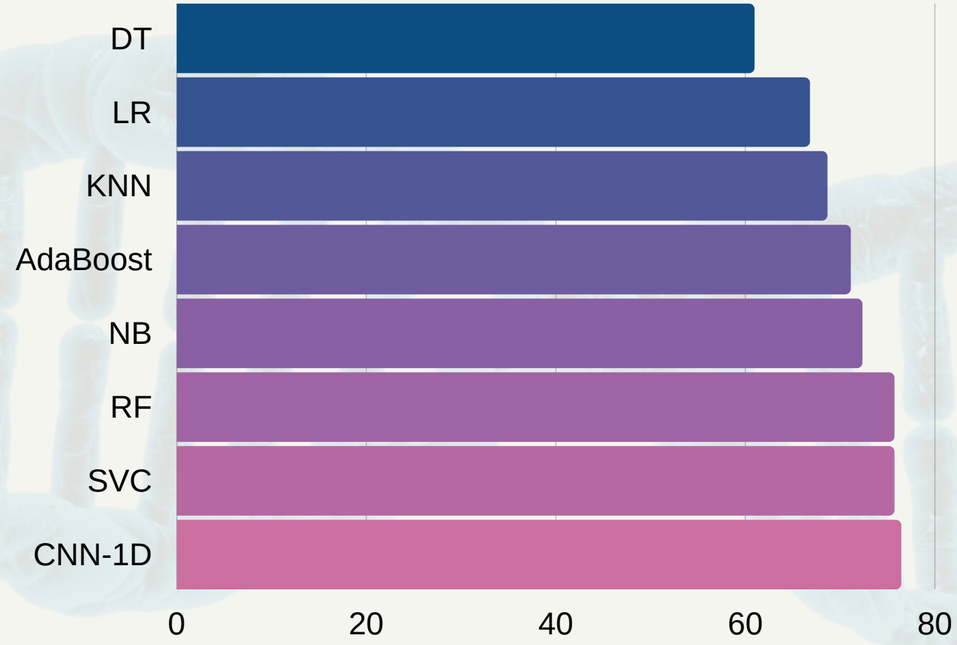


Figure 5: Algorithms accuracy

Methodology - (CNN-1D)

CNN-1D model description:

To reduce overfitting we've used Dropout and Regularization.
We've also used Batch Normalizer to normalize our data on every step.
Finally, **Softmax** function gives us result on probability format.

Parameters	Values
Filters	16
Kernel Size	4
Kernel Regularizer l2	0.01
Dropout (rate)	0.80
Dense Layer 1(unit)	16
Dense Layer 2(unit)	8
Dense Layer 3(unit)	2

Figure 6: Used Parameters (CNN-1D)

Result Analysis - (Algorithms)

DT was the poor performer in terms of MCC and accuracy. But SVC and RF performs well in terms of Sp and Sn respectively. These two algorithms also performs well in terms of MCC.

However, **we selected CNN-1D as the best performing classifier** and suitable for our method as its accuracy was highest of 76.41% and superior to RF in terms of MCC and Pc which are 0.53 and 85.55% respectively.

Classifier	Sn(%)	Sp(%)	MCC	Pc(%)	Acc(%)
SVC	57.05	91.48	0.52	85.00	75.69
NB	60.40	82.39	0.44	74.39	72.31
RF	57.72	90.91	0.52	84.31	75.69
AdaBoost	59.06	81.25	0.41	72.73	71.08
LR	61.74	71.02	0.33	64.34	66.77
KNN	55.70	79.55	0.36	69.75	68.62
DT	58.39	63.07	0.21	57.24	60.92
CNN-1D	60.04	89.97	0.53	85.55	76.41

Figure 7: Comparison of performances of different classification algorithms

Result Analysis - (CNN-1D ROC Curve)

In Machine Learning, performance measurement is an essential task. So when it comes to a classification problem, we can count on an AUC - ROC Curve. When we need to check or visualize the performance of the multi - class classification problem, we use AUC (Area Under The Curve) ROC (Receiver Operating Characteristics) curve. It is one of the most important evaluation metrics for checking any classification model's performance. Figure-8 shows the performance of our CNN-1D model.

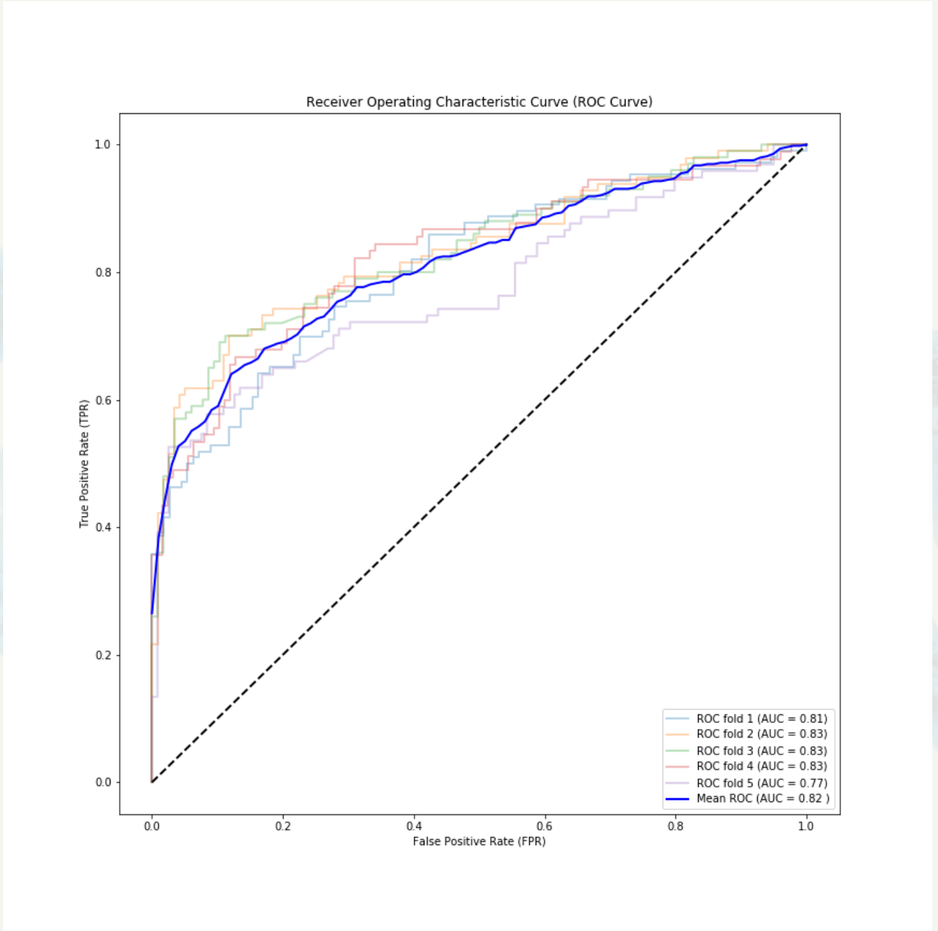


Figure 8: CNN-1D ROC curve

Result Analysis - (State-of-the-art-predictors)

Now, we will compare the performance of iRecSpot-CNN-1D with other state-of-the-art predictors who used the same datasets. We have chosen six previous predictors for the purpose of comparison. They are: iRSpot-TNCPseAAC, iRSpot-PseDNC, IDQD, iRSpot-ADPM, iRSpot-SF and iRSpot-EF.

Results in terms of Sensitivity (Sn), Specificity (Sp), Accuracy (Acc) and MCC are reported in Figure 9.

From the results shown in Figure 9, it is evident that our proposed method **iRecSpot-CNN-1D could not outperforms** all these methods in terms of all the evaluation metrics considered for the experiments and, classification methods and feature selection approaches employed by other state-of-the-art pre-dictors is also given in 9.

Methods	Sn(%)	Sp(%)	MCC	Acc(%)
iRSpot-TNCPseAAC	76.56	70.99	0.4737	73.52
iRSpot-PseDNC	71.75	85.84	0.5830	79.30
IDQD	79.52	81.82	0.6160	80.77
iRSpot-ADPM	77.19	90.73	0.6905	84.57
iRSpot-SF	84.57	75.76	0.6941	84.58
iRSpot-EF	94.35	95.80	0.9037	95.14
iRSpot-CNN-1D	**60.04**	**89.97**	**0.5345**	**76.41**

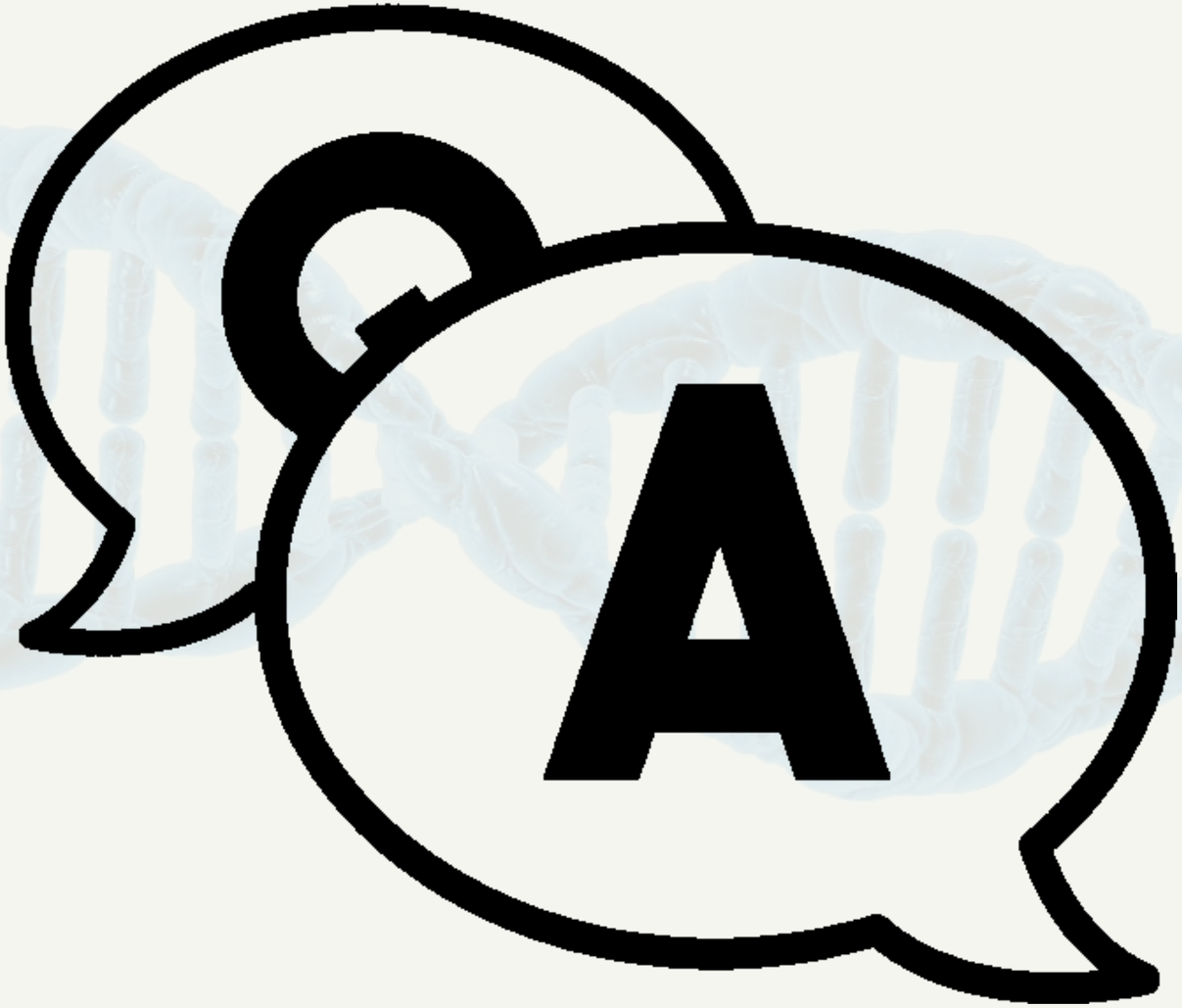
Figure 9: Comparison of performances of iRecSpot-CNN-1D with other state-of-the-art predictors

Conclusion

In this project, we have worked on a predictor, iRecSpot-CNN-1D to efficiently identify recombination spots employing a set of novel features in combination with a coherent feature selection approach. In total, seven hundred sixty features were generated by the three feature generation method. Our features As iRecSpot-CNN-1D trying to accomplish to identify recombination spots better than any other existing methods, it may play a significant role in genetics and cell biology.

Furthermore, the simplicity and the efficiency of the method is very promising as a bioinformatics approach to reveal the mechanism of recombination and genome variation. We are trying more optimization approaches to generate features.

Q&A



Thank You