**DNA Classification**

Sambhav Gulla - 202018018, Shreya Patel - 202018028

Jay Shah - 202018051, Namra Jain - 202018059

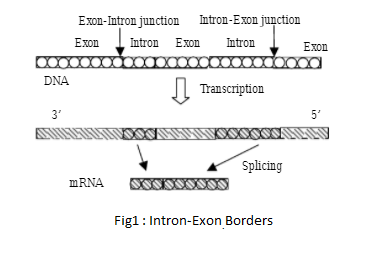
**Abstract** - **Sequence classification is to predict the type of DNA sequence based on the similarity of its structure or function, predict the sequence function and the relationship between other sequences, predict the class, and assist in identifying genes in DNA molecules. In this paper, Four Machine Learning algorithms namely, Decision Tree, Random Forest, AdaBoost, and SVM are used for the task of DNA classification. We have used Two feature extraction techniques namely One-Hot Encoding and N-gram. The performance of machine learning techniques in order of increasing accuracy is as follows SVM > Adaboost > Random Forest > Decision Tree suggesting that SVM is one of the most dependable algorithms for splice junction detection.**

**Keywords**: **DNA classification; bioinformatics; machine learning; feature extraction**

1. Introduction

DNA is the biochemical material that carries hereditary characteristics from parents to offspring is contained in a sequence of chemicals known as deoxyribonucleic acid (DNA). DNA analysis has become a vital interest in computational biology. A gene consists of a continuous sequence of DNA that is needed to produce a particular protein of mRNA. The process by which the DNA gives rise to a protein is called gene expression. Due to the splicing, a DNA sequence consists of alternating segments of exon and intron, where an exon is a nucleotide sequence that is expressed or translated into protein, and an intron is an intervening sequence that is transcribed into RNA but later eliminated from the transcript by splicing its adjacent exons. The splice junction identifies the point where the splicing takes place, i.e., it is the meeting point of intron(non-coding regions) and exon(coding regions) are joined together.

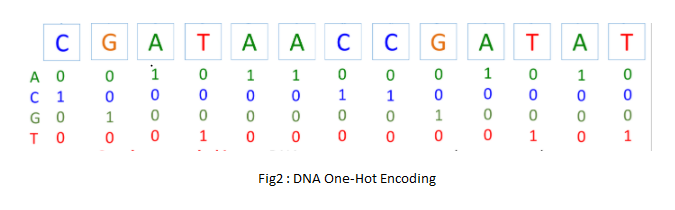
Splicing is one of the post-transcriptional modiﬁcations which prepare the mature m-RNA. When proteins are synthesized, some superfluous DNA is removed from sites known as splice sites. Our DNA Classification mainly consists of three classes IE, EI and N which are the boundary side between Intron-Exons, Exons-Intron and neither respectively. Introns are used to splice in and Exons are used to splice out. Pre-mRNA is made when exons are spliced out. IE borders are known as “acceptors” while EI borders are known as “donors”. The DNA sequence in our dataset is 60 fields long. Each of these fields is filled by one of A, G, T, C.



1. Feature Extraction Methods

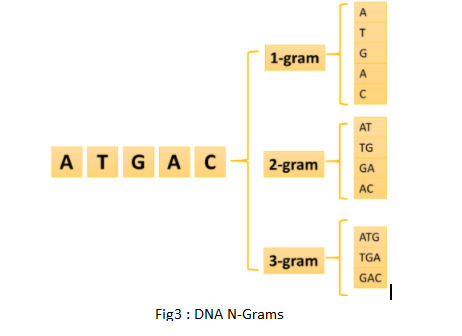
# A. One Hot Encoding :

We can not use categorical data directly with machine learning algorithms. With One hot encoding, we can represent categorical data as binary vectors. Instead, we use character-level one-hot encoding to represent the categorical data(for us DNA). Each base is encoded as a vector of all zeros except one in a specific position, A, T, C, and G are encoded as 0001, 1000, 0010, and 0100 respectively.



## B. N-Grams :

The method we use here is simple and easy. The long DNA sequence is taken and is broken down into k-mer overlapping words For example if we use words of length 4, ‘ATGAC’ becomes: ‘ATG’, ‘TGA’, ‘GAC’. Therefore our DNA sequence is transformed into 3 3-mer words, here we are using 3-mer "words”. The word length and amount of overlap need to be determined empirically for any given application. In genetic science, we call this method of breaking down DNA sequences k-mer counting.



1. Machine Learning Algorithms

# **A.** Decision Tree :

Decision trees are constructed via an algorithmic approach that identifies ways to split a data set based on different conditions. It is one of the most widely used and practical methods for supervised learning. Decision Trees are a non-parametric supervised learning method used for both classification and regression tasks. Tree models where the target variable can take a discrete set of values are called classification trees.

**B.** Random Forest :

The random forest model trains multiple decision trees at training times on different data and outputs the class output basis on the average prediction of the individual trees. There is a result at each node, and the data are classified into different child nodes. The child nodes show the final results. Two hyperparameters needed to be tuned in our implementation. One of them is the number of estimators. It controls the number of trees that should be created during the experiment. The other one is the maximum depth of each tree. This value should neither be too large nor too small. Larger tree numbers may be the reason for the overfitting, while lower depths could lead to underfitting.

**C.** AdaBoost :

It makes *n* number of decision trees during the training period of data. As the first decision tree/model is made, the record which is incorrectly classified during the first model is given more priority. Only these records are sent as input for the second model. The process will go on until we specify the number of base learners we want to create. Remember, the repetition of records is allowed with all boosting techniques.

**D.**  SVM :

The SVM aims to find the hyperplane which maximizes the margin separating the two classes. Powerful non-linear SVM models can be trained if kernel functions are appropriately used. Kernel functions create new feature vectors that usually have more dimensions than the original input. The SVM finds the new hyperplane, which is linear in the new feature space. However, in the original feature space, the separation will be non-linear if a non-linear kernel is used

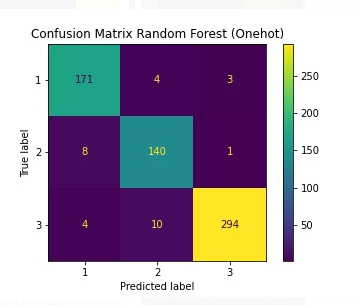
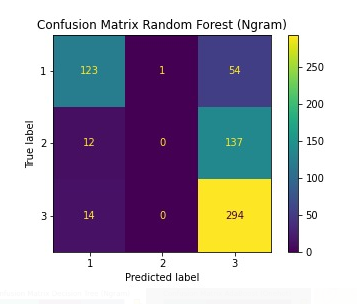
1. Results :

Decision Tree :

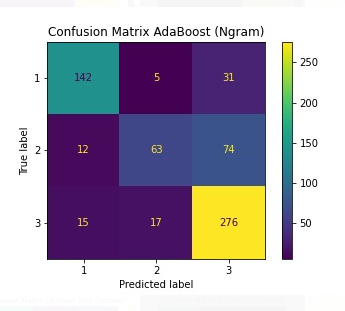
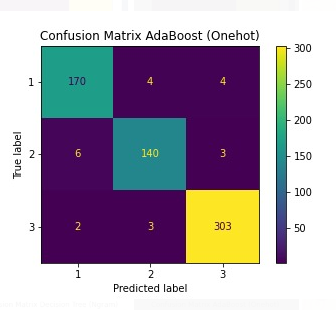
## 

## 

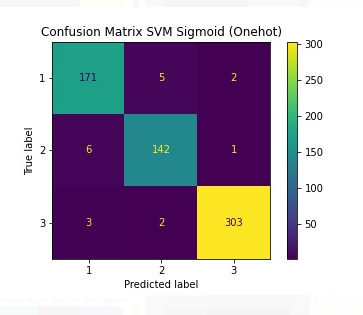
## Random Forest :

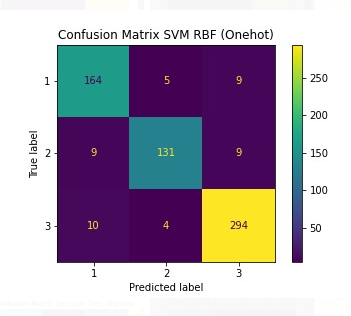
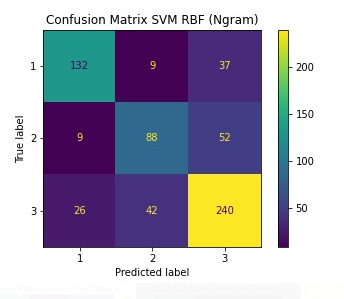


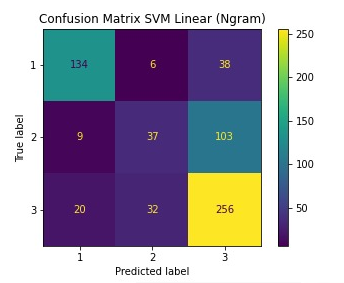
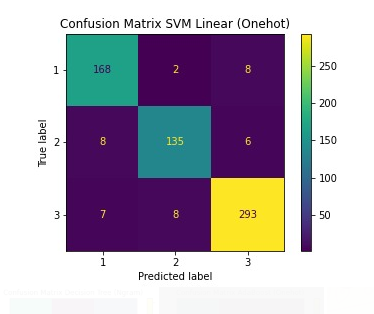
## AdaBoost :



## SVM :





**Accuracy:** All results presented are the mean accuracy over ten folds of cross-validation. the results of all six machine learning algorithms using the random DNA sequence The feature extraction method is presented in the Table below

| Model | One-hot Accuracy | Ngram Accuracy |
| --- | --- | --- |
| Decision Tree | 75.9 | 57.67 |
| Random Forest | 93.1 | 63.81 |
| Adaboost | 95.6 | 74.6 |
| SVM(Linear) | 93.5 | 68.2 |
| SVM(RBF) | 92.04 | 69.6 |
| SVM(Sigmoid) | 96.3 | 72 |

V. Conclusion :

In this paper, we compared two different feature extraction methods: OneHot encoding and the 3-gram method. We then passed these features to 4 different machine learning classifiers: Decision Tree, Random Forest, Adaboost, and SVM with 3 different kernel functions. Machine learning methods when fed with one hot encoding feature gave higher accuracy than that with the N-gram method. The SVM classifier had the highest accuracy of 96.3 percent, while the Decision tree classifier had the lowest accuracy of 75 percent. Other classifiers' accuracies were above 90 percent. For the Ngram features, the highest accuracy was achieved by the Adaboost classifier, while the lowest accuracy was reported by the decision tree classifier.

VI. References

<https://www.researchgate.net/publication/334283027_Splice_Junction_Prediction_In_DNA_Sequence_Using_Multilayered_RNN_Model>

<https://www.researchgate.net/publication/28357925_Recognition_of_DNA_Splice_Junction_via_Machine_Learning_Approaches>

<https://www.researchgate.net/publication/224055329_Splice_junction_classification_problems_for_DNA_sequences_representation_issues>

<https://apps.dtic.mil/dtic/tr/fulltext/u2/a409780.pdf>

<https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-7-S5-S15>

<https://genomebiology.biomedcentral.com/articles/10.1186/gb-2014-15-1-r19>