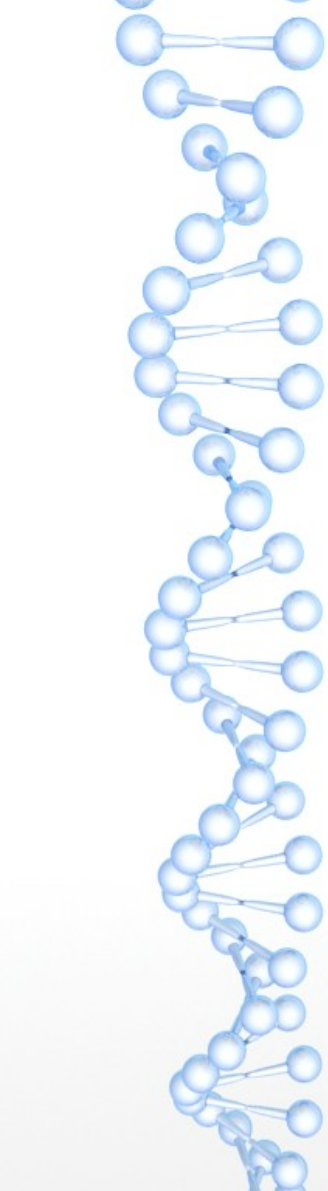


A deep Learning Approch to DNA Sequence Classification

Purpose: To extract meaningful features from raw data and use these features for the classification of gene sequences.

In this work is present a deep learning neural network for the classification of DNA sequences, based on the spectral representation of the sequence.

The work is tested on a dataset of 16S genes and its performance in terms of accuracy and F1-score.

It is shown that the presented convolutional network is "better" for this type of task than other learning models.

A deep Learning Approch to DNA Sequence Classification

Dataset: 16S, taxonomy.csv

Steps:

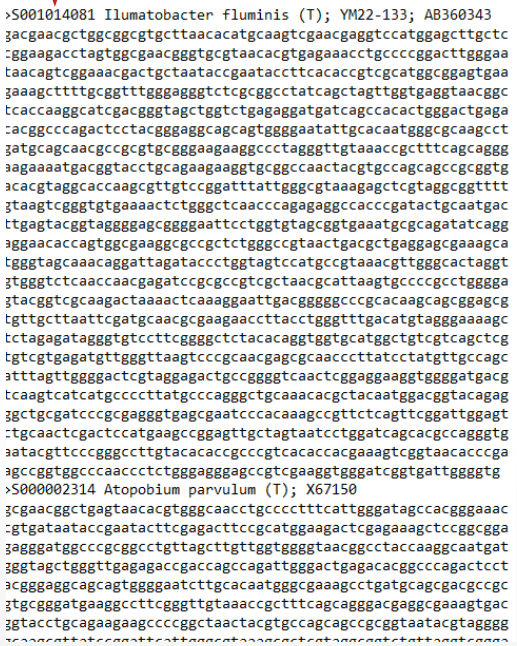
- Data preprocessing : Sequence Cleaning and Spectral Representation in k-mers
- Use of CNN (Convolutional Neural Network) based on LeNet5
- Comparison of the results with the results obtained by other models (SVM, NB, RF)

Datasets

Taxonomy.csv

Sequence	PHYLUM	CLASS	ORDER	FAMILY	GENUS
S001014081	Actinobacteria	Actinobacteria	Acidimicrobiales	Acidimicrobidae_incertae_sedis	Ilumatobacter
S000002314	Actinobacteria	Actinobacteria	Coriobacteriales	Coriobacteriaceae	Atopobium
S000004268	Actinobacteria	Actinobacteria	Coriobacteriales	Coriobacteriaceae	Atopobium
S000130242	Actinobacteria	Actinobacteria	Coriobacteriales	Coriobacteriaceae	Atopobium
S000390775	Actinobacteria	Actinobacteria	Coriobacteriales	Coriobacteriaceae	Atopobium
S000414609	Actinobacteria	Actinobacteria	Coriobacteriales	Coriobacteriaceae	Atopobium
S001100366	Actinobacteria	Actinobacteria	Coriobacteriales	Coriobacteriaceae	Enterorhabdus
S000013627	Actinobacteria	Actinobacteria	Coriobacteriales	Coriobacteriaceae	Coriobacterium
S000734935	Actinobacteria	Actinobacteria	Coriobacteriales	Coriobacteriaceae	Asaccharobacter
S001168715	Actinobacteria	Actinobacteria	Coriobacteriales	Coriobacteriaceae	Gordonibacter

16s.fasta

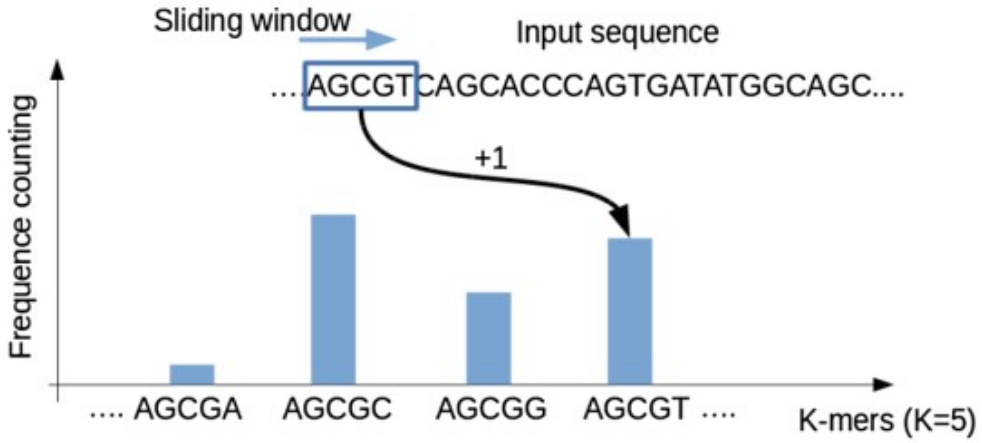


# of seq = 3000
Length of sequences > 1200bp

In total we have the sequences divided as follows:

Three main bacteria phyla	Number of categories for each taxa				
	Phylum	Class	Order	Family	Genus
Actinobacteria	1	1	3	12	79
Firmicutes	1	2	3	19	110
Proteobacteria	1	2	13	34	204

Spectral Representation



CNN
(Convolutional Neural Network)

Network architecture:

1° Layer

Filters = 10
kernel_size=5

2° Layer

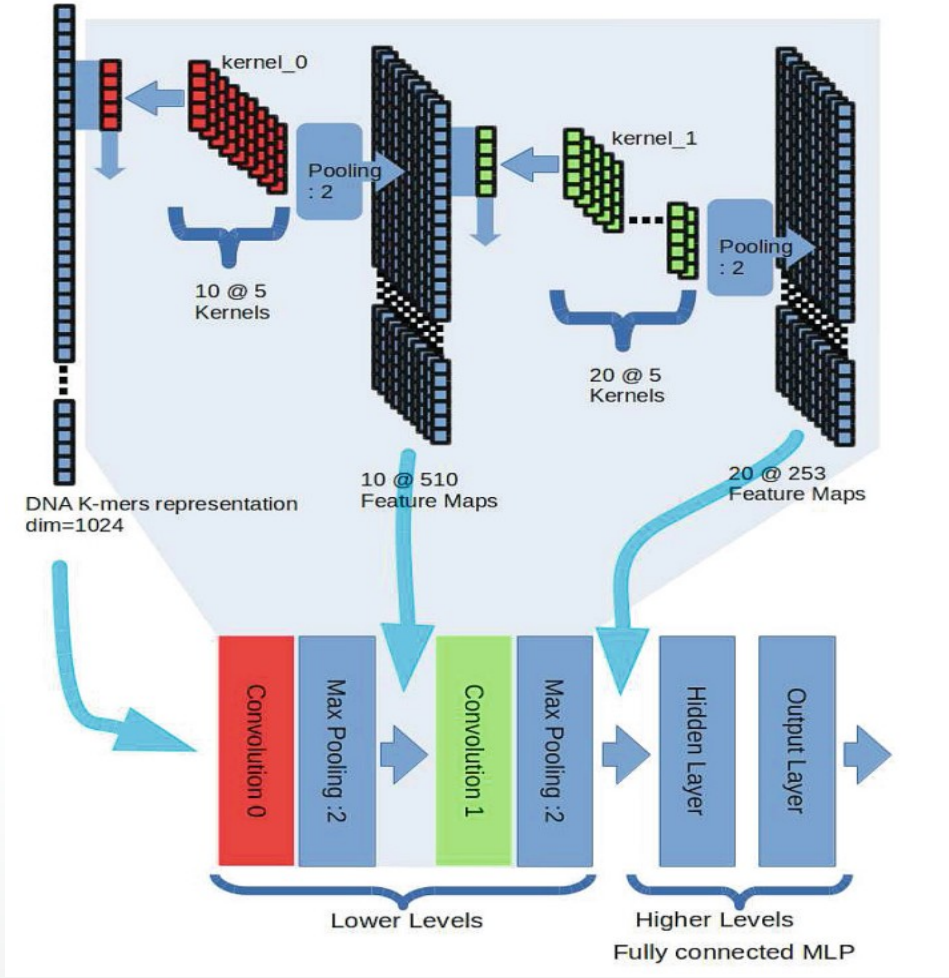
Filters = 20
kernel_size=5

Output of each level:

$$dim(q) = \frac{dim(x) - (dim(w) - 1)}{\text{size of max-pooling}}$$

Hidden Layer: 500 units

Output Layer: 1 units for each class



1.Input Sequences

The length of the input sequences considered is:

Stage 1: around 1400bp for sequence (full)

Stage 2: 500bp for sequence (short)

3.Metrics

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

$$F1 = \frac{2TP}{2TP + FP + FN} = 2 * \frac{precision * recall}{precision + recall}$$

2.Models

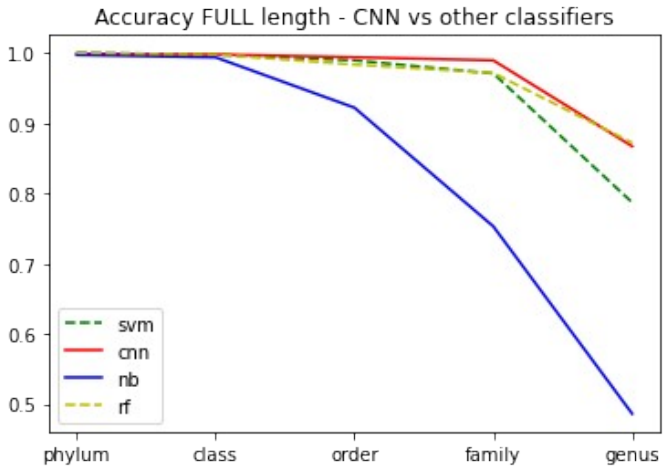
CNN

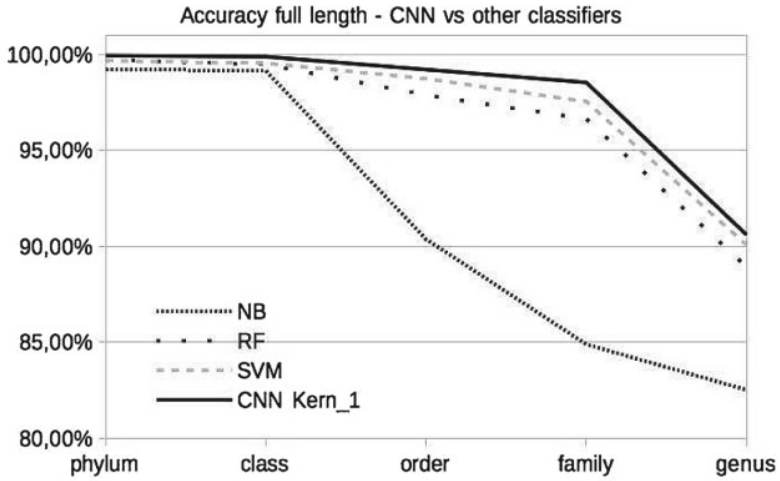
SVM Support Vector Machine con Kernel Gaussian Radial Basis

Naive Bayesian

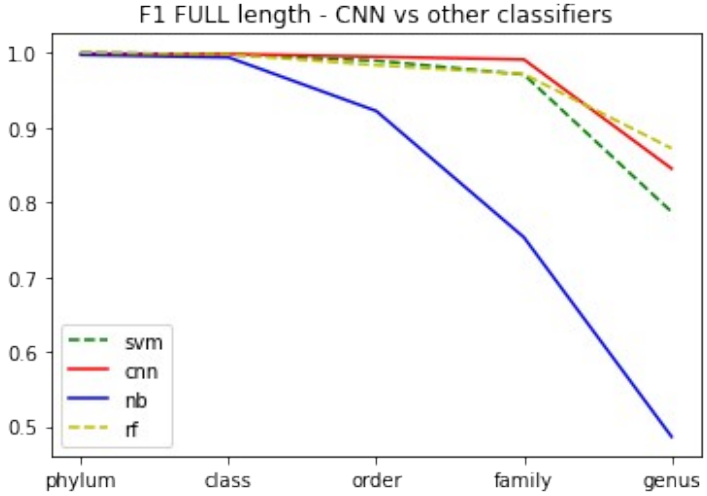
Random Forest

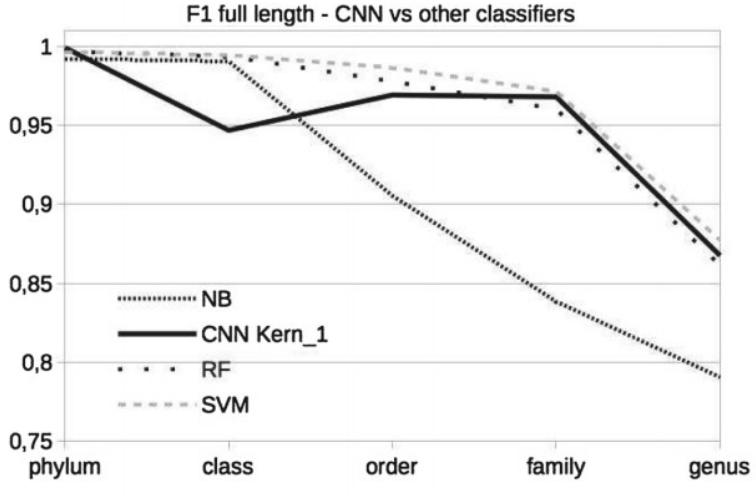
Results obtained on full-length Accurancy (>1200bp) max-pooling stride=2



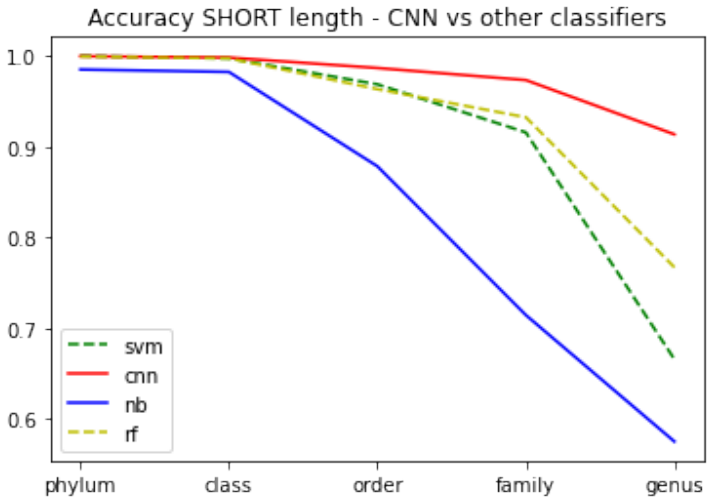


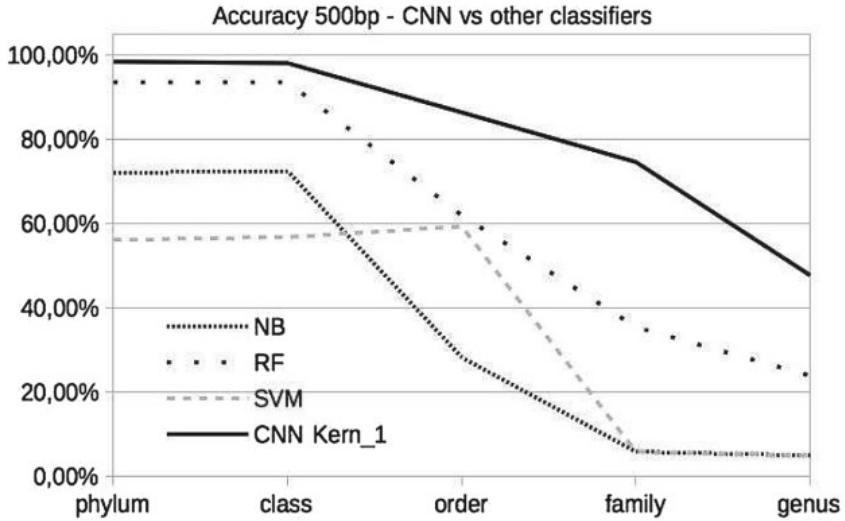
Results obtained on full-length F1 (>1200bp) max-pooling stride=2



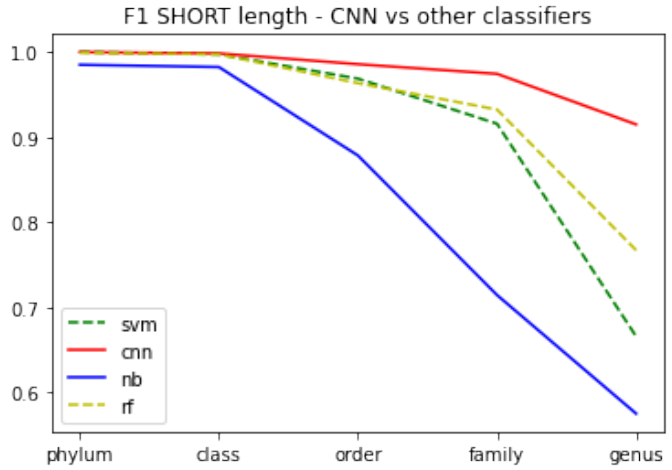


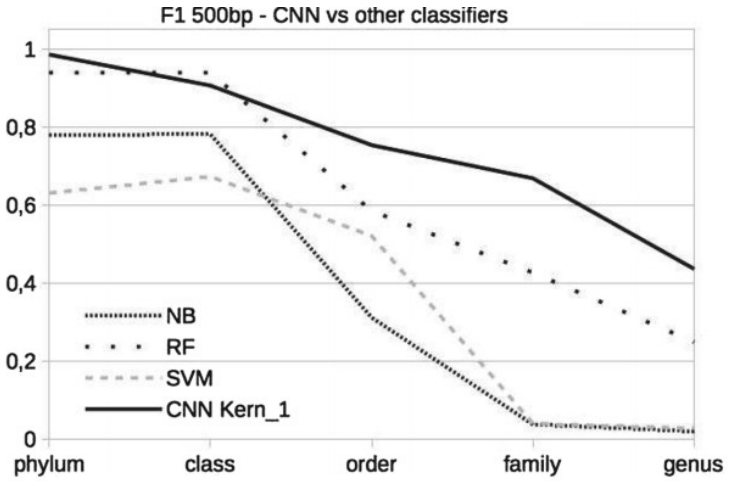
Results obtained on short length Accuracy (500bp) max pooling stride=2



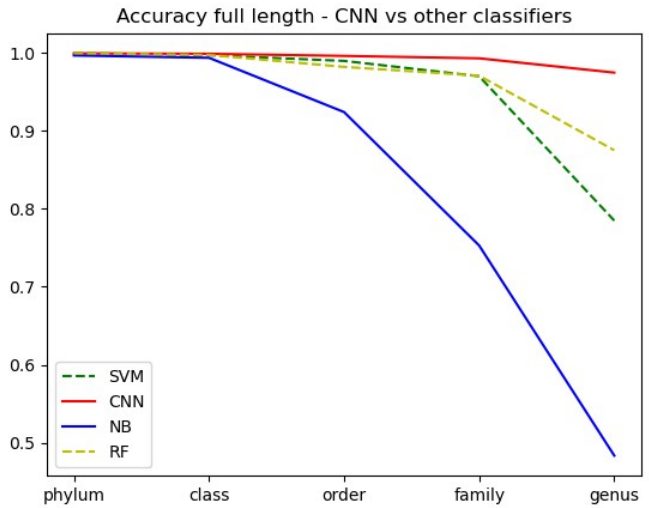


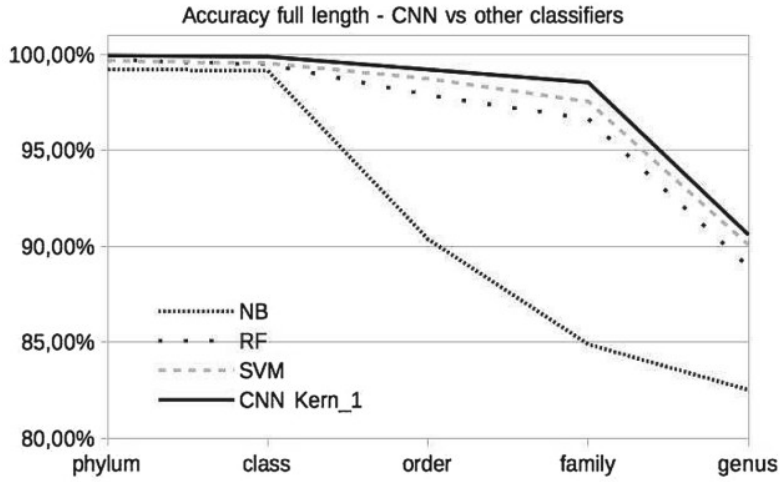
Results obtained on short length F1 (500bp) max pooling stride=2



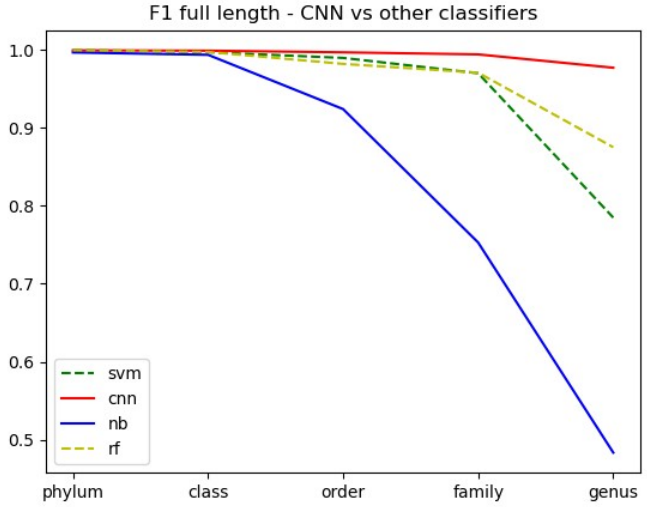


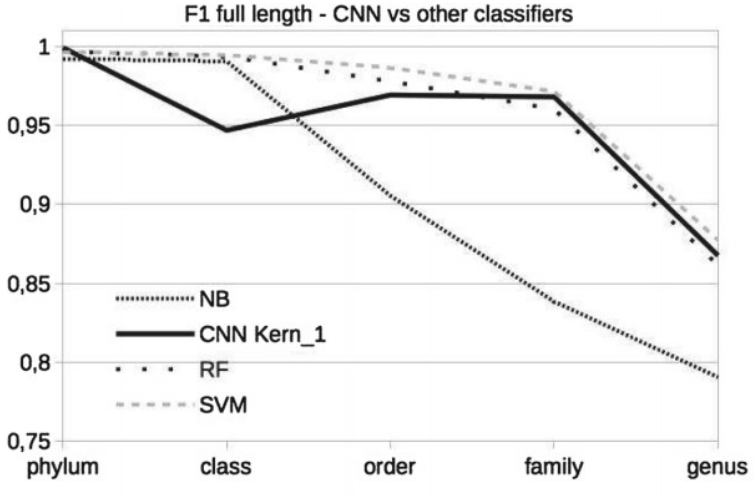
Results obtained on full-length Accurancy (>1200bp)



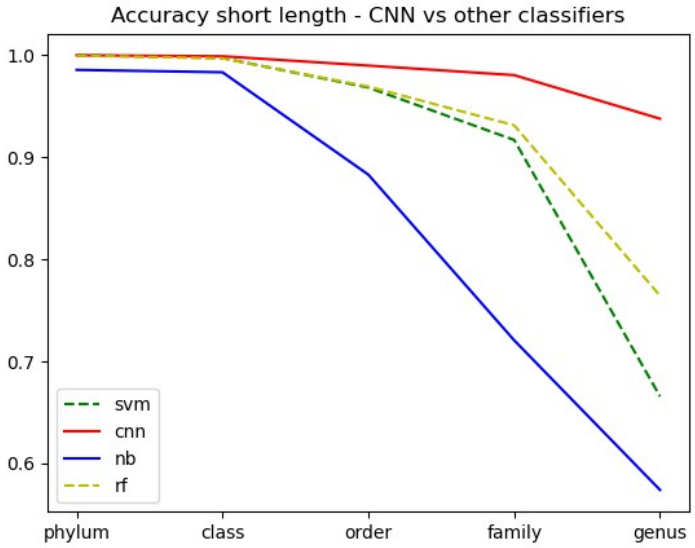


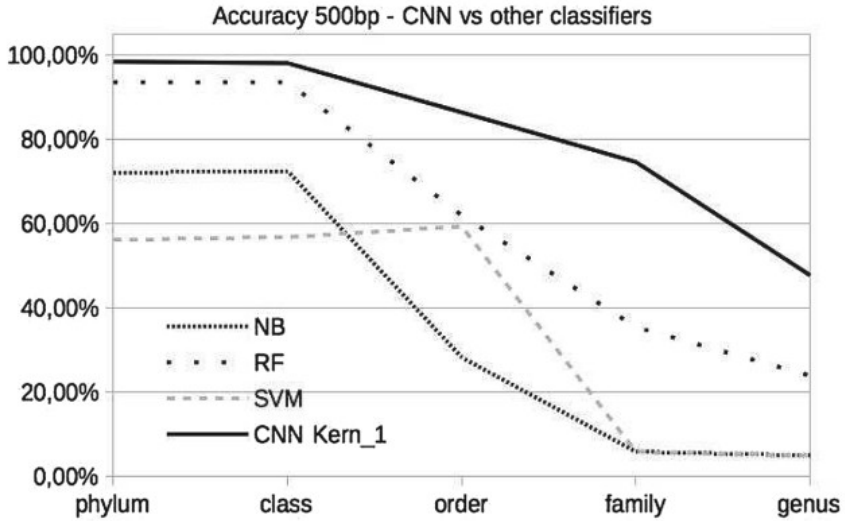
Results obtained on full-length F1 (>1200bp)





Results obtained on short length Accuracy (500bp)





Results obtained on short length F1 (500bp)