CMSC 435: INTRODUCTION TO DATA SCIENCE
FINAL PROJECT REPORT
12/2/2025

TEAM NAME: Anthony_Lightfoot_Johnson

TEAM MEMBERS:
David Anthony
Jaidon Lightfoot
Benjamin Johnson

Design Description:

This project aimed to build a system that classifies protein sequences as DNA, RNA, DRNA, or nonDRNA. Because the data consisted only of amino-acid sequences, we focused on converting each sequence into meaningful numerical features and testing several models in RapidMiner. After comparing different feature sets and classifiers, we selected the best-performing design and used it to generate predictions for the blind test dataset.

Design 1 was the initial baseline model. We created basic amino-acid composition features, including counts and frequencies of each residue and broad biochemical groups such as hydrophobic, acidic, and polar amino acids. We trained a default Decision Tree on this initial feature set to establish a starting point for comparison.

Design 2 expanded the feature set by adding additional biochemical indicators, including motif counts, ratios, and more detailed frequency-based features. Although the feature space was larger, we still used a Decision Tree classifier so we could directly evaluate the impact of adding new attributes. This design showed improvement over the baseline but still struggled with the smaller binding classes.

Design 3 combined the complete feature set from the earlier designs and used the stored normalization model to keep preprocessing consistent. We trained a tuned Random Forest classifier, which handled the expanded feature set effectively and achieved the strongest overall accuracy and class-level performance. Because it consistently outperformed the earlier designs, we selected this model for generating the blind test predictions

Results:

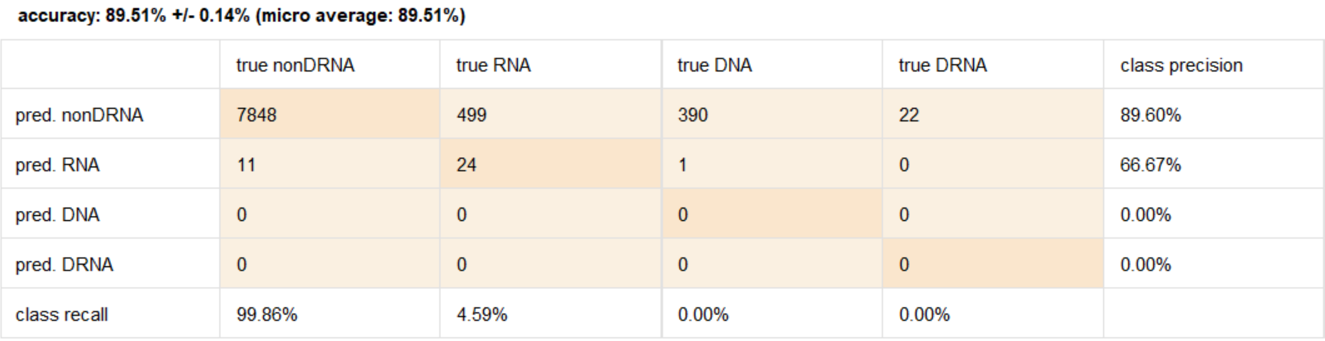
Table 1. Summary of results based on the 5-fold cross-validation on the training dataset.

Outcome	Quality measure	Baseline result	Design 1	Design 2	Design 3	Best Design
DNA	*Sensitivity*	6.9	0.0	11.8	5.6	5.6
	Specificity	99.3	100.0	98.4	99.9	99.9
	Accuracy	95.2	95.6	94.5	95.8	95.8
	MCC	**0.132**	**0.0**	**0.108**	**0.217**	**0.217**
RNA	*Sensitivity*	39.6	4.6	29.8	38.2	38.2
	Specificity	98.9	99.9	98.3	99.5	99.5
	Accuracy	95.3	94.1	94.1	96.0	96.0
	MCC	**0.501**	**0.075**	**0.390**	**0.542**	**0.542**
DRNA	*Sensitivity*	4.5	0.0	0	9.1	9.1
	Specificity	100.0	100.0	99.95	100.0	100.0
	Accuracy	99.7	99.8	99.7	99.8	99.8
	MCC	**0.122**	**0.0**	**0**	**0.301**	**0.301**
nonDRNA	*Sensitivity*	98.6	99.9	97.3	99.5	99.5
	Specificity	29.8	2.7	28.5	25.1	25.1
	Accuracy	91.3	89.5	90.0	91.6	91.6
	MCC	**0.428**	**0.185**	**.472**	**0.439**	**0.439**
averageMCC		**0.296**	**0.065**	**0.243**	**0.375**	**0.375**
accuracy4labels		90.8	89.5	89.2	91.5	91.5

Confusion matrix that corresponds to the baseline result

		predicted			
		DNA	RNA	DRNA	nonDRNA
actual	DNA	27	20	0	344
	RNA	21	207	1	294
	DRNA	0	2	1	19
	nonDRNA	36	71	1	7751

Design 1 Matrix:

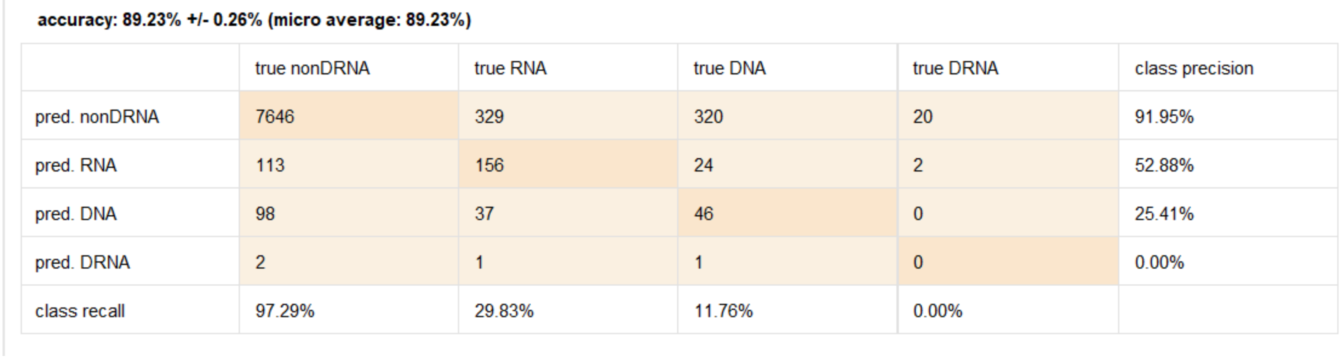


accuracy: 89.51% +/- 0.14% (micro average: 89.51%)

	true nonDRNA	true RNA	true DNA	true DRNA	class precision
pred. nonDRNA	7848	499	390	22	89.60%
pred. RNA	11	24	1	0	66.67%
pred. DNA	0	0	0	0	0.00%
pred. DRNA	0	0	0	0	0.00%
class recall	99.86%	4.59%	0.00%	0.00%	

- Features
    - Sequence length
    - Amino Acid composition
- Preprocessing
    - Read CSV
    - Rename columns
    - Generate IDs
    - Generate attributes
    - Set Role
    - 5-fold cross-validation
- Classification algorithm
    - Decision tree

Design 2 Matrix:



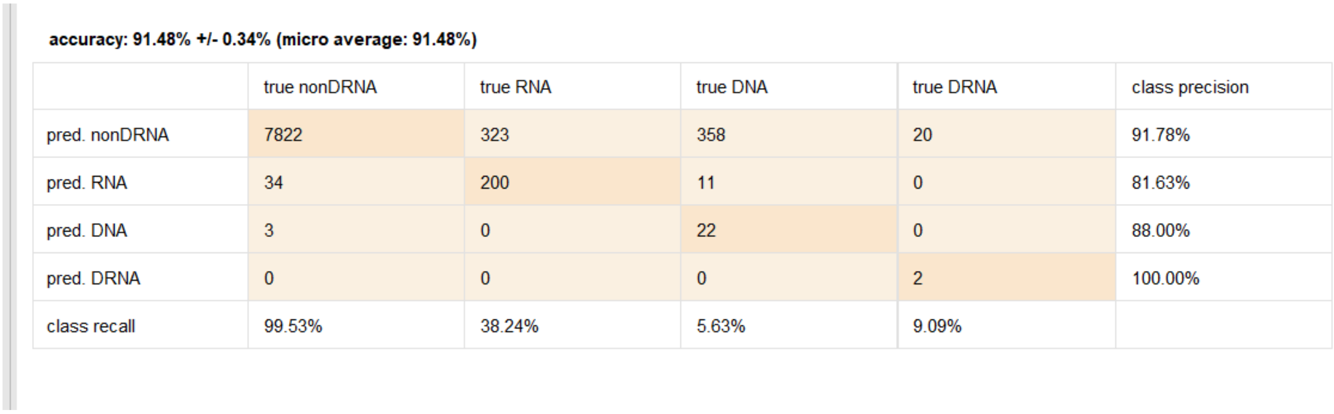
accuracy: 89.23% +/- 0.26% (micro average: 89.23%)

	true nonDRNA	true RNA	true DNA	true DRNA	class precision
pred. nonDRNA	7646	329	320	20	91.95%
pred. RNA	113	156	24	2	52.88%
pred. DNA	98	37	46	0	25.41%
pred. DRNA	2	1	1	0	0.00%
class recall	97.29%	29.83%	11.76%	0.00%	

- Features
  - Sequence length
  - Amino Acid composition
  - Inverse length
  - Amino acid count
  - Amino acid frequency
  - Count based on chemical group (basic, acidic, aromatic, hydrophobic, polar)
  - Frequency based on chemical group (basic, acidic, aromatic, hydrophobic, polar)
  -
- Preprocessing
  - Read CSV
  - Rename columns
  - Generate IDs
  - Generate attributes
  - Normalize
  - Set Role
  - 5-fold cross-validation
- Classification algorithm
  - Decision tree
    - criterion = gain_ratio
    - Maximal depth from 10 to 50
    - Confidence from .1 to 0.05
    - Minimal gain from .1 to .0
    - Minimal leaf size from 2 to 1
    - Minimal size for split from 4 to 2

- Observations
  - DNA sensitivity increased
  - RNA sensitivity increased
  - DNA & RNA MCC improved
  - nonDRNA MCC also improved

Design 3 Matrix (using random forest):



accuracy: 91.48% +/- 0.34% (micro average: 91.48%)

	true nonDRNA	true RNA	true DNA	true DRNA	class precision
pred. nonDRNA	7822	323	358	20	91.78%
pred. RNA	34	200	11	0	81.63%
pred. DNA	3	0	22	0	88.00%
pred. DRNA	0	0	0	2	100.00%
class recall	99.53%	38.24%	5.63%	9.09%	

- Features
  - Sequence length
  - Amino Acid composition
  - Inverse length
  - Amino acid count
  - Amino acid frequency
  - Count based on chemical group (basic, acidic, aromatic, hydrophobic, polar)
  - Frequency based on chemical group (basic, acidic, aromatic, hydrophobic, polar)
- Preprocessing
  - Read CSV
  - Rename columns
  - Generate IDs
  - Generate attributes
  - Normalize
  - Set Role
  - 5-fold cross-validation
- Classification algorithm
  - Random Forest
    - Number of trees from 100 to 200
    - Maximal depth from 10 to 50
    - Voting strategy from confidence vote to majority vote

- Observations
  - DNA sensitivity increased
  - RNA sensitivity increased
  - DNA & RNA MCC improved
  - nonDRNA MCC also improved

Test Predictions: **Total: 8794**

	DNA	RNA	DRNA	nonDRNA
Total:	32	254	0	8508
Ratio:	0.36%	2.89%	0%	96.75%

Conclusion:

This project demonstrated how both feature quality and model choice directly influence performance on the four-class protein interaction task. Our early designs, which relied on basic sequence composition features and Decision Trees, performed noticeably worse than the baseline provided in Table 1, especially in sensitivity for the minority classes. After expanding the feature set and switching to a Random Forest in Design 3, we achieved stronger accuracy, higher MCC values, and improved class-level sensitivity, particularly for DNA and RNA. Although DRNA remained difficult to classify because of its very small sample size, our best model still outperformed the baseline on key metrics and showed more balanced behavior across classes.

The iterative design process was useful for understanding how sequence derived features affect model performance. Decision Trees were easy to interpret but struggled with large feature sets and noisy patterns, while the Random Forest handled the expanded feature space more effectively and reduced overfitting. A limitation of our approach is the continued difficulty in detecting rare DRNA proteins, but the advantage is that our final model offers consistent improvements across most metrics and represents a meaningful advancement over the baseline. Overall, the project showed how important it is to build the right features, choose the right model settings, and evaluate results carefully when working with protein sequence data.