CMSC 435 Project

Fall 2025
(Group work; 27 pts total)

The project asks your project group to conceptualize, design, evaluate, and compare models for the prediction of proteins that interact with DNA and RNA using a provided real-life dataset. Your model must classify a given protein sequence into one of four outcomes, i.e., interacts with DNA (DNA), interacts with RNA (RNA), interacts with both DNA and RNA (DRNA), and does not interact with DNA or RNA (nonDRNA). Although each group will solve the same task, the corresponding designs must be unique, i.e., collaboration between groups is not allowed. The underlying objective is to create a simple prototype of a solution and gradually improve it to make it more accurate.

This is a group project, where groups must include 2 or 3 students. Students will select their teammates, i.e., create groups. You can use the "Discussion" section on the class site in Canvas if you need to find teammates. The instructor reserves the right to randomly assign students to groups in case they do not complete the team selection by the deadline. Note that while we allow smaller, 2-person groups, such groups are expected to deliver the same projects as the 3-person groups. Each group must complete the team contract as its first deliverable; see the "Deadlines and Delivery" section of this document for the deadline and details. Please also note that October 17 is the deadline for the mid-semester grades, which may affect the class withdrawals and commitment to the class.

Data

Two datasets are provided:

- *sequences_training.txt* (*training dataset*) that includes 391 DNA proteins, 523 RNA proteins, 22 DRNA proteins, and 7859 nonDRNA proteins, for a total of 8795 proteins. This dataset is already available in Canvas.
- *sequences_test.txt* (*blind test dataset*) that includes 8794 proteins, with similar proportions between the four classes of proteins. This is an independent test set, which means that the entire design procedure (including feature generation, feature selection, parameterization, and selection of classifiers, etc.) should be completed using only the training dataset. The test dataset should be used to evaluate your system only once. This dataset will be posted on the class website 2 days before the project submission deadline, and it will **not** include the annotation of the outcomes. You will have to predict the outcomes, and the instructor will process and assess these predictions.

The training dataset is provided in the comma-separated format, where each protein is represented by:

- the amino acid sequence
- the class label encoded as DNA, RNA, DRNA, and nonDRNA

The test dataset will be the same format as the training dataset, except that the outcomes will not be provided.

Evaluation of Predictions

Your group is required to perform the **5-fold cross-validation** when using the *training dataset*. This cross-validation divides the training dataset into 5 random, equal-sized subsets, where one subset is used to test the prediction model and the remaining four to train/develop the prediction model; this is repeated 5 times, each time using a different subset as the test set. Consequently, this test results in predicting every sequence in the training dataset. This test procedure is supported by the AI Studio (RapidMiner) and other popular data science software/libraries.

For each of the four outcomes, your group will convert the dataset into a binary problem, i.e., a given outcome (positive outcome) vs. all other outcomes (negative outcomes). For example, all proteins that are labeled as DNA will be considered positive, and the remaining proteins (RNA, DRNA, and nonDRNA) will be considered negative. Next, for each of the four outcomes, you will compute the following measures:

Sensitivity = SENS = 100*TP / (TP + FN)

Specificity = SPEC = 100*TN / (TN + FP)
Accuracy = 100* (TP+TN) / (TP+FP+TN+FN)
MCC = (TP*TN – FP*FN) / sqrt[(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)]

where TP is the number of true positives (correctly predicted positive outcomes), FP denotes false positives (negative outcomes that were predicted as positives), TN denotes true negatives (correctly predicted negative outcomes), and FN stands for false negatives (positive outcomes that were predicted as negatives). You will also compute:

averageMCC = $(MCC_{DNA} + MCC_{RNA} + MCC_{DRNA} + MCC_{nonDRNA})/4$
accuracy4labels = $100*TP_{all}$ / (number of all protein in the dataset)

where MCC_{DNA}, MCC_{RNA}, MCC_{DRNA}, and $MCC_{nonDRNA}$ denote the MCC values when using the DNA, RNA, DRNA, and nonDRNA outcomes as the positives, and TP_{all} is the number of correctly predicted outcomes (DNA proteins predicted as DNA proteins, RNA proteins predicted as RNA proteins, etc.). These measures can be computed based on the confusion matrix. You should **round the values** to one digit after the decimal point when reporting the accuracy, sensitivity, and specificity, and to three digits after the decimal point when reporting MCC (see Table 1 for examples). **Your report must include the confusion matrix for your final/best solution**.

Your group must also provide and **summarize predictions on the *blind test dataset***. To do that, you will compute your model using the entire training dataset (using the same design, i.e., features, values of parameters, etc., as in your best 5-fold cross-validation result), and you will use this model to predict sequences from the blind test dataset. **In your report, you must discuss the corresponding results on both the training and the blind test dataset**; on the blind test dataset, you can summarize your results by explaining and comparing how many proteins were predicted with a given outcome.

<u>*Design*</u>

Your group should **design** the model to maximize its predictive performance, **evaluated based on averageMCC using the 5-fold cross-validation on the training dataset**. The design may consider:

- Use of different features to encode the input protein sequence. The data mining algorithms require a rectangular dataset with a fixed size and structure of the feature vector for each object (protein). Thus, you will need to convert the input protein sequences (that have variable length) into a fixed set of (numerical) features. Lecture set 7 includes a few suggestions.
- Selection of a subset of the input features. This could potentially speed up the computation of the model, remove weak/noisy features, and reduce overfitting. Feel free to combine the results of multiple feature selection methods.
- Selection of a classification algorithm that you will use to compute your model from among many algorithms that are available in the AI Studio (RapidMiner) or other software that you will use.
- Parametrization of the selected classification algorithm(s). This involves identifying and setting values of their key parameters.
- Building a system with multiple models that are used together. For instance, you could use multiple models that predict all 4 classes and combine their results together to generate one prediction. Check the methods in the AI Studio at Operators → Modeling → Predictive → Ensembles.

IMPORTANT NOTE 1: Ensure that your team performs all design activities (e.g., feature selection, selection and parametrization of the classification algorithms, etc.) using the 5-fold cross-validation on the training dataset. Otherwise, you could overfit this dataset, and your results on the blind test dataset will suffer.

IMPORTANT NOTE 2: If you will perform sampling, ensure that you do not sample the test folds in the cross-validation, i.e., the cross-validation results must be done on the five test folds that collectively include the 8795 proteins.

IMPORTANT NOTE 3: Your team's design should be done **incrementally**. Start with a simple initial solution (complete the entire design, prediction, and prediction assessment process) and gradually make your design more sophisticated with the goal of improving the predictive performance. The progress report checkpoint should be based on a simple initial solution. In your final report, you should clearly indicate **the best set of results**, which must be selected based on the cross-validation results on the training dataset. Moreover, these results should be compared with your intermediate results (earlier/simpler designs, other alternatives, etc.) and with baseline results shown in Table 1, in order to justify your design choices. **In your final report, provide your results by adding them to Table 1.** This will make it easy to compare the different alternatives. **Clearly indicate which result is the best/final**. You should explain how you made decisions that led you in the direction of redesigning/improving your model. You also should provide a convincing argument why and how your method is good/competitive in comparison to the **real baseline result** listed in Table 1.

Table 1. Predictive results based on the 5-fold cross-validation on the training dataset (this table is available in Canvas).

Outcome	Quality measure	Baseline result	Design 1	Design 2	Design 3	Best Design
DNA	*Sensitivity*	6.9				
	Specificity	99.3				
	Accuracy	95.2				
	MCC	**0.132**				
RNA	*Sensitivity*	39.6				
	Specificity	98.9				
	Accuracy	95.3				
	MCC	**0.501**				
DRNA	*Sensitivity*	4.5				
	Specificity	100.0				
	Accuracy	99.7				
	MCC	**0.122**				
nonDRNA	*Sensitivity*	98.6				
	Specificity	29.8				
	Accuracy	91.3				
	MCC	**0.428**				
averageMCC		**0.296**				
accuracy4labels		90.8				

Deliverables

Each group shall provide the following **four deliverables**:

1. **Team project contracts** (one per team), filled in and signed.
2. **Progress report** (PDF file for parts a and b; dataset file for part c) that consists of:
   a) **Description of the first attempt to make predictions**. You should use short bullet points to list the features that you generated from the input sequences; list major data processing steps that you used to prepare the data; and name the classification algorithm that you used. This is supposed to be an initial attempt, so we expect to see simple models and just a few bullet points.
   b) **4x4 confusion matrix and the four MCC values** ($\text{MCC}_{\text{DNA}} + \text{MCC}_{\text{RNA}} + \text{MCC}_{\text{DRNA}} + \text{MCC}_{\text{nonDRNA}}$) that the above model has produced.
   c) **Training dataset file** in txt/csv format. This file is the rectangular dataset with a fixed set of features for each object (protein), where the last (right-most) feature is the class. This is supposed to be an initial version of the dataset, and so we expect to see a small and simple feature set.
3. **Final Report** (PDF file) that consists of:
   - **Cover page** that gives the class number and title, date of your submission, name of your group, and names of all team members.
   - **Description of the design of the prediction system**. You should briefly **explain** the features that you generated from the input sequences; **how** and **which** features were selected; **which** classification algorithms and their parameters you tried and **why,** and **which** you have chosen; and **which** other design options you considered and applied.

   - **Results** (see *Evaluation of Predictions* section). You must **organize the results in a table** using the format of Table 1. Using this format, compare your best cross-validation results with the results from earlier/alternative designs and with the results shown in Table 1. Include a confusion matrix for your best solution. Summarize predictions for the *blind test dataset*.
   - **Conclusions**. This is a **very important part** of your report. You should **comment on the quality of your results** and **compare** them against the baseline results from Table 1. Also, describe **your experience** in this project, and explain the **advantages** and **disadvantages** of your method and why you think your results are good or bad, in comparison with the other results from Table 1.
4. **Predictions on the *blind test* dataset**. These predictions should be submitted via email (see *Deadline and Delivery* below) as a text file named with the name of your group, where each row provides a prediction for a given "blind" protein. The format should be as follows:

   DNA
   DNA
   RNA
   nonDRNA
   …

   where DNA, RNA, DRNA and nonDRNA are the predicted outcomes for the protein from the same row in the *sequences_test.txt* file. The instructor will use these results to evaluate your method on the blind test dataset against the true classes, and these results will be forwarded to you as part of the evaluation of your project.

Marking

The evaluations of the progress report, project report, and predictions constitute **27% of the final mark from the course,** and it will consist of the following four parts:

1. 7% for the quality of the progress report
2. 8% for the quality of the final report
3. 6% for the quality of the design of the prediction method from the final report
4. 6% for the quality of the predictions measured using the 5-fold cross-validation on the training dataset from the final report and on the blind test dataset

IMPORTANT NOTE 4: For item 4, the *averageMCC* is the main predictive quality measure that will be used to evaluate submitted solutions, but the conclusions **must discuss** the other quality indices as well. **Bonuses of 5%, 4%, 3%, 2%, and 1% will be awarded to the project submissions that secure the highest, the second highest, the third highest, the fourth highest, and the fifth highest value of *averageMCC* on the blind test dataset.** In case of a tie, the winner will be decided based on the higher value of the *accuracy* on the blind test dataset.

IMPORTANT NOTE 5: For item 4, MCCs that are high(er) relative to other submissions or to the baseline in Table 1 are not necessary to receive a full project mark. The key is to show substantial progress from the initial solution. You should show and discuss how your best design is better when compared to your own alternative solutions and explain the advantages compared to the baseline results in Table 1.

Deadlines and Delivery

- Filled in and signed **team project contracts** (one per team) must be returned to the instructor **by email** (to lkurgan@vcu.edu) **before 11:59 pm** on **October 23 (Thursday), 2025**. Remember to **copy your submission email to all members of the team**. The project contract form is available in Canvas and should be editable using a PDF reader application.
- Submission of the progress report deliverables (PDF file accompanied by the file with the training dataset; two files in total; one submission per group) to the instructor and TA **by email** (lkurgan@vcu.edu and yuj11@vcu.edu) is due **before 11:59 pm** on **November 6 (Thursday), 2025**. Remember to **copy your submission email to all members of the team**. This PDF file should include the bullet-point description,

confusion matrix, and MCC values. Make sure to clearly identify the team name/number and members of the submitting group. Late submission penalties apply.

- Submission of the project deliverables (final report PDF file and the test predictions; two files in total; one submission per team) to the instructor and TA **by email** (lkurgan@vcu.edu and yuj11@vcu.edu) is due **before 11:59 pm** on **December 2 (Tuesday), 2025**. Remember to **copy your submission email to all members of the team**. Late submission penalties apply.

Final Notes

- Read this document carefully and, in particular, pay attention to the **text in bold**.
- While we recommend the use of the AI Studio (RapidMiner), you can complete this project using any other data science software or language. Just make sure that you will be ready to make predictions on the blind test dataset using your selected software.
- Do not cheat (e.g., do not inflate or "tweak" the results). It is better to report honest results than to get caught cheating. In the latter case, you risk receiving 0 marks for the project.
- Your team may be asked to demonstrate how the prediction works in case the reported results are irregular. Thus, make sure to retain your software at least until the time of the final exam.
- Always copy the email communications to yourself so you can prove that it was sent.
- Contact the instructor immediately if problems occur.