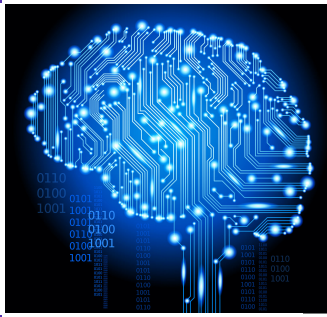
# Machine learning applications to DNA subsequence and restriction site analysis

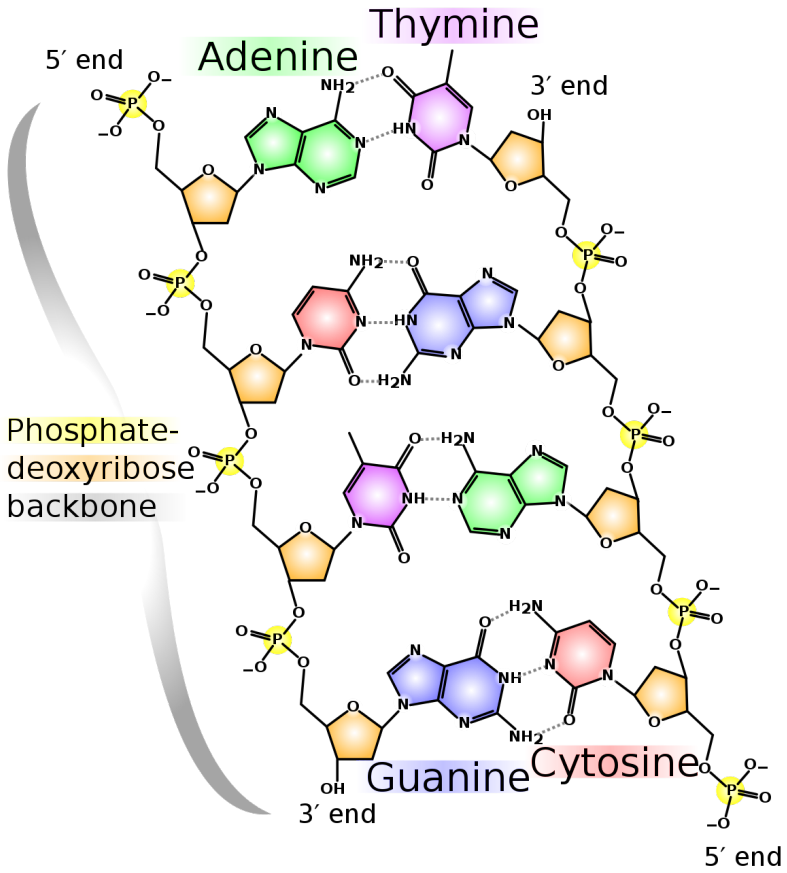
Ethan Jacob Moyer



Drexel Distributed, Intelligent, Scalable COmputing (**DISCO**) Lab

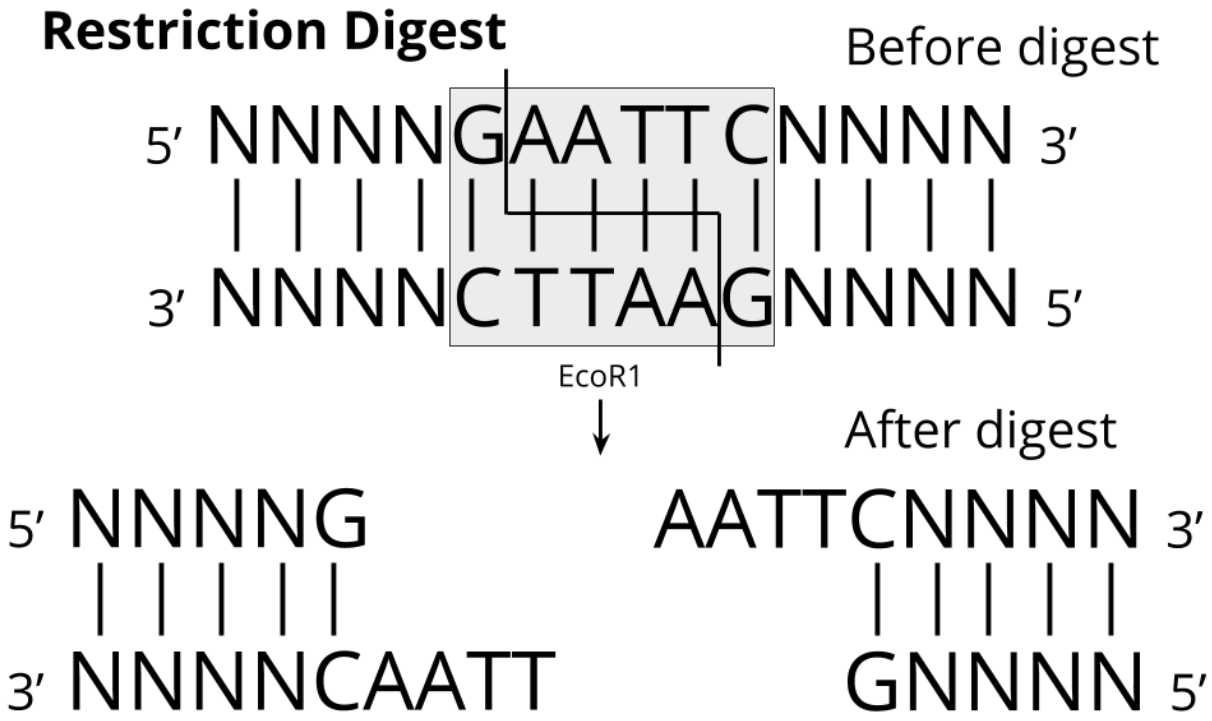
Drexel UNIVERSITY

# Introduction



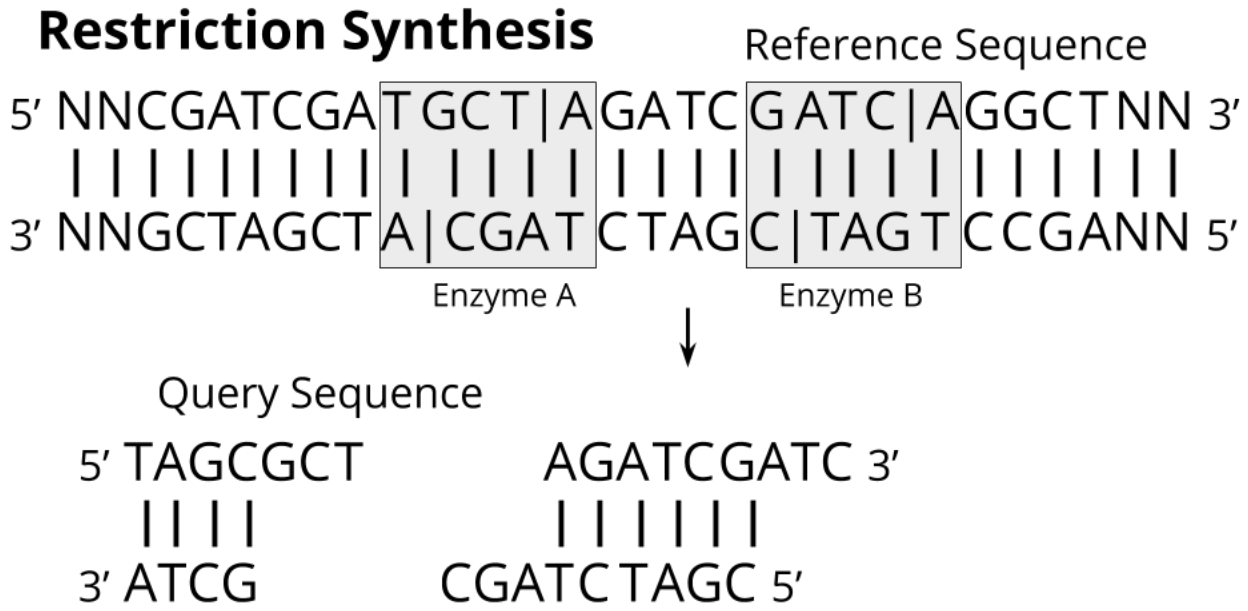

Madeleine Price Ball. 2020

# Introduction




**Restriction digest:** lab procedure where enzyme cleaves DNA.

# Introduction




**Restriction synthesis:** novel catabolic DNA synthesis that relies on iterative restriction enzyme digest and sticky end ligation of fragments.

# Feature Set

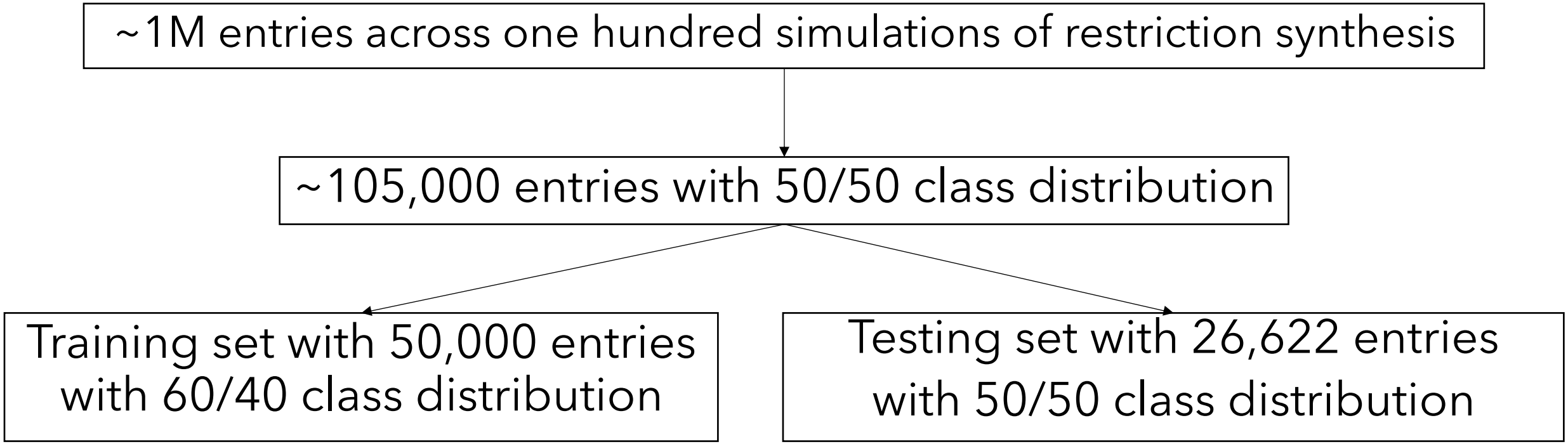
- SEQ & LEN
- C & EZY
- $p(x, L) = \frac{occurrence\ of\ x}{L}$
- $r_1(L) = 1 - \sum_{i \epsilon \alpha}(\frac{1}{4} - p(i, L))^2$ , $where\ \alpha = \{A, T, C, G\}$
- $r_2(L) = \sum_{i=b}^{p} \frac{\sum_{j=0}^{n-i} r_1(i) * \frac{i}{n}}{\sum_{k=b}^{p} \frac{k}{n}}$, $where\ b = 4,\ p = 16, and\ j\ is\ the\ starting\ index$

**Reference**

.....

**Subsequence**

# Data Set


~1M entries across one hundred simulations of restriction synthesis
~105,000 entries with 50/50 class distribution
Training set with 50,000 entries with 60/40 class distribution
Testing set with 26,622 entries with 50/50 class distribution

# Methods

- Support Vector Machine (SVM)
- Random Forest
- Convolution Neural Network (CNN)

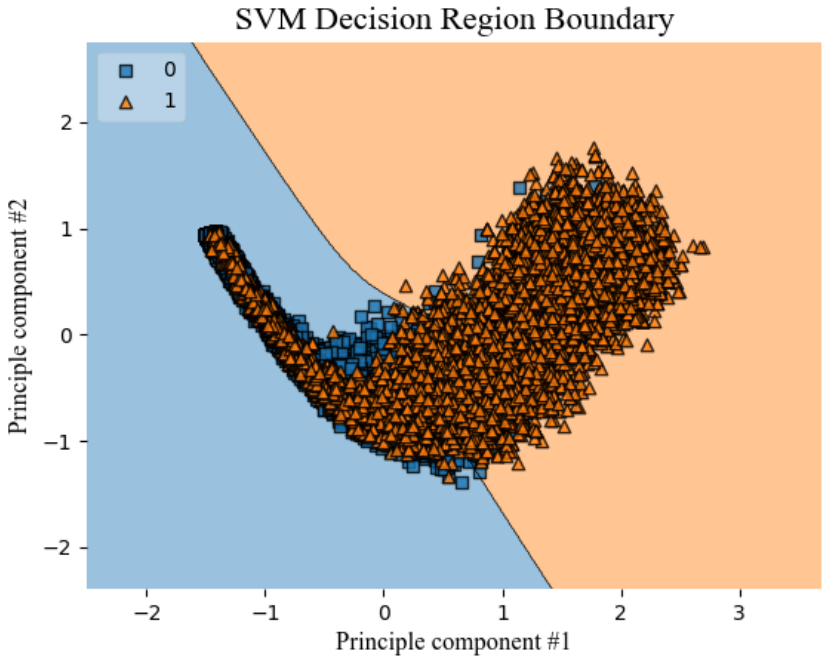
# Methods - Metrics

- Sensitivity = $\frac{true\ postive}{true\ positive + false\ negative}$
- Specificity = $\frac{true\ negative}{true\ negative + false\ positive}$
- False negative rate (FNR) = 1 - Sensitivity
- False positive rate (FPR) = 1 - Specificity

# Methods - Support Vector Machine (SVM)

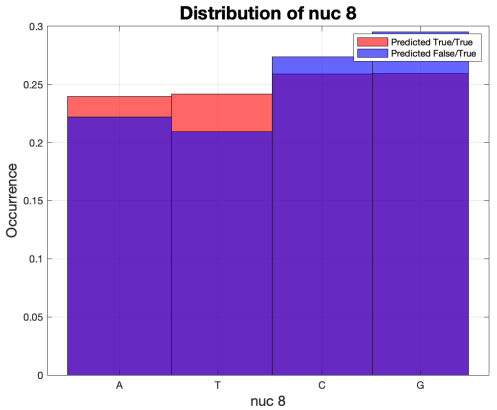
- **SVM:** a supervised model that creates non-probabilistic hyperplanes by maximizing distances between classes

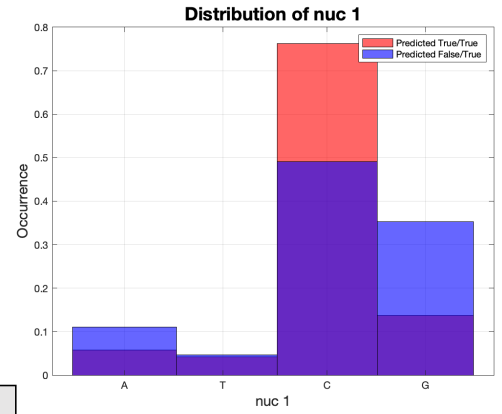
- **Principle Component Analysis (PCA):** an unsupervised model that reduces the feature space into  a set number of components



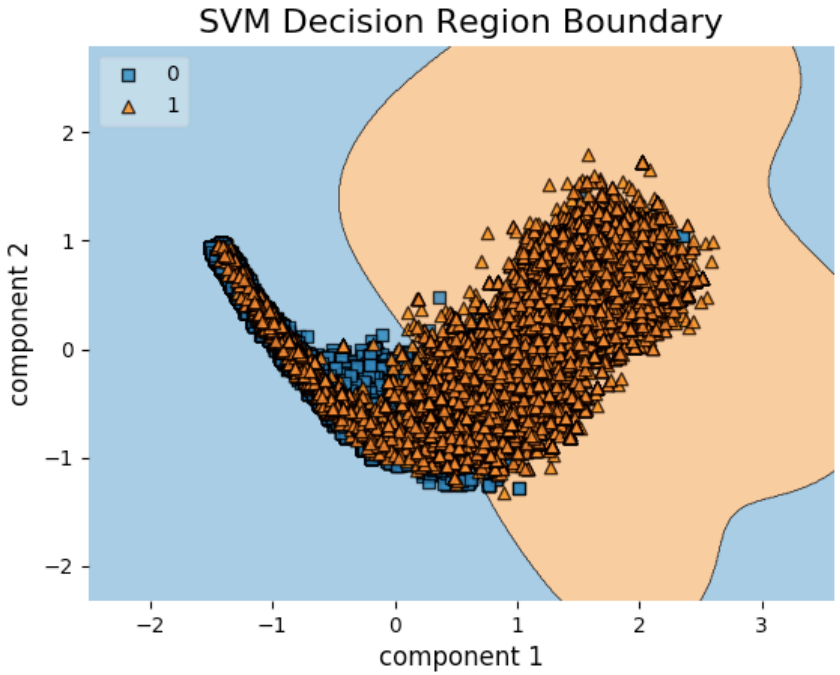

| | Actual True | Actual False |
|---|---|---|
| Predicted True | 12632 | 3009 |
| Predicted False | 678 | 10304 |

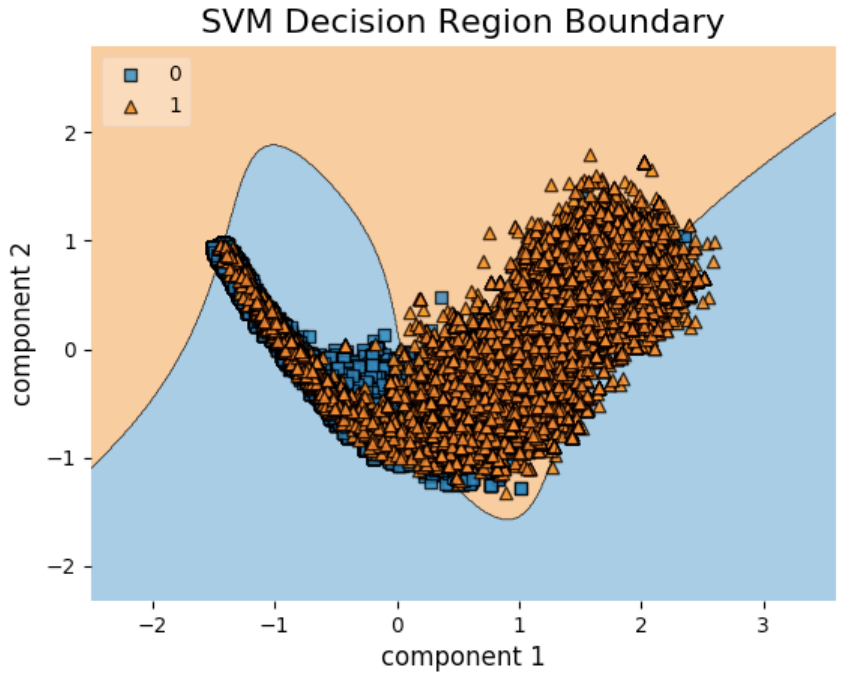
| Classifier | Sensitivity (%) | Specificity (%) | FNR (%) | FPR (%) |
|---|---|---|---|---|
| 2 PCs polymetric kernel | 94.9 | 77.4 | 5.1 | 22.6 |



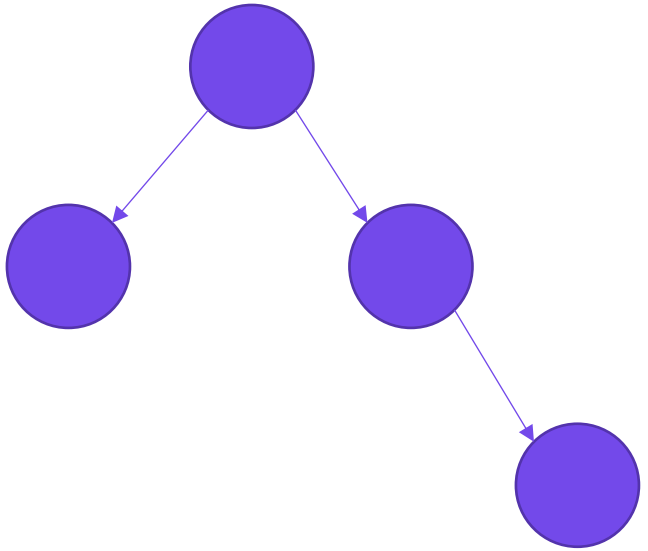
# Methods - Support Vector Machine (SVM)




| Classifier | Sensitivity (%) | Specificity (%) | FNR (%) | FPR (%) |
|---|---|---|---|---|
| Non-PCA linear kernel | 89.3 | 83.6 | 10.7 | 16.4 |
| 3 PCs linear kernel | 92.5 | 81.3 | 7.5 | 18.7 |
| 3 PCs polymetric kernel | 92.7 | 81.7 | 7.3 | 18.3 |
| 2 PCs RBF kernel | 93.1 | 81.7 | 6.9 | 18.3 |
| 2 PCs sigmoid kernel | 75.4 | 75.5 | 24.6 | 24.5 |

# Methods - Random Forest



- **Random Forest:** an ensemble learning model that uses a set number of decision trees on a subset of the data.
- **Pros:** Average out overfitting
- Tested n = 10, 20, 30, 40, 50

| | Actual True | Actual False |
|---|---|---|
| Predicted True | 12336 | 1898 |
| Predicted False | 974 | 11415 |

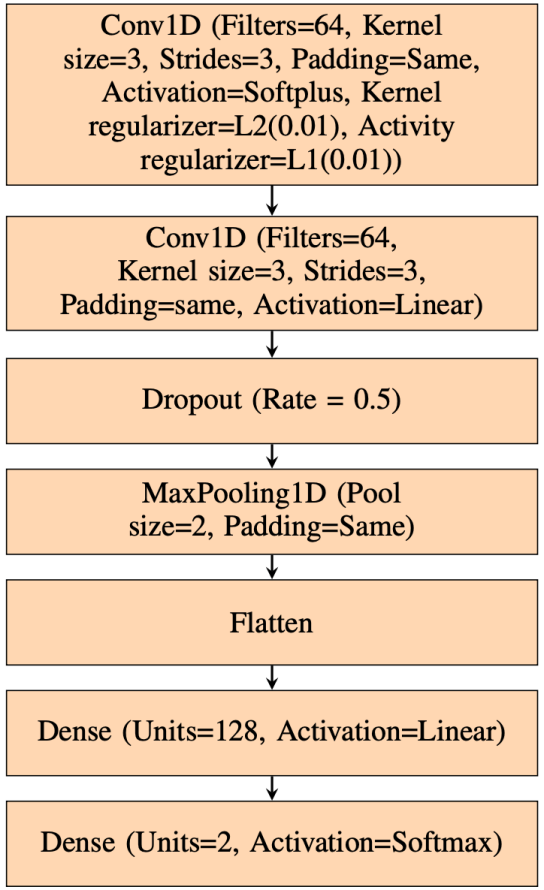
| Classifier | Sensitivity (%) | Specificity (%) | FNR (%) | FPR (%) |
|---|---|---|---|---|
| Random Forest (n=30) | 92.7 | 85.7 | 7.3 | 14.3 |

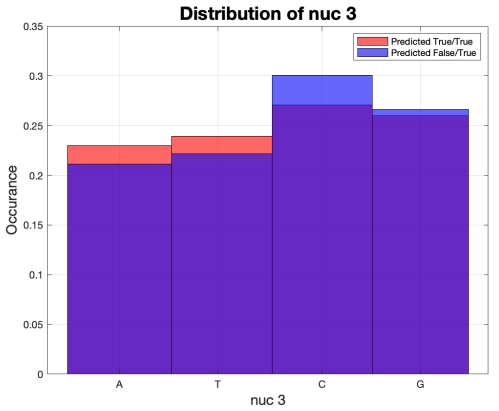
# Methods - Convolution Neural Network (CNN)

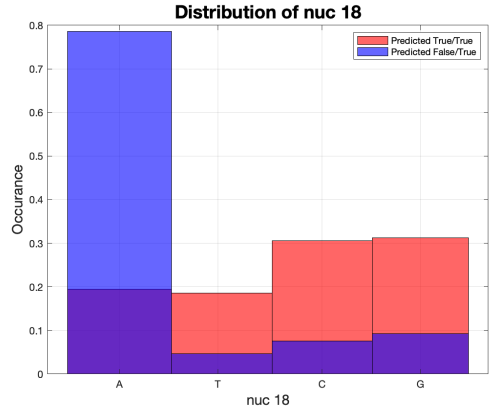
- **CNN:** deep learning model that use the convolution operation in at least one of their layers

- Produced variable results, so metrics were averaged over 100 classifications

| Classifier | Sensitivity (%) | Specificity (%) | FNR (%) | FPR (%) |
|---|---|---|---|---|
| CNN | 91.4 | 82.4 | 8.6 | 17.6 |



# Discussion

- High false positive rate--as opposed to high false negative rate--is preferred
  - False positive: including a subsequence when it should not have been included
  - False negative: not including a subsequence when it should have been included

- Recursive nature of the problem

- Sensitivity: SVM (94.9%) > Random Forest (92.7%) > CNN (91.4%)

# Thank you

Acknowledgements

This work is supported by the National Science Foundation Award CCF-1937419 (RTML: Small: Design of System Software to Facilitate Real-Time Neuromorphic Computing