Machine Learning Uncovers CCM Isoforms as Transcription Factors

Jacob Croft[1,] *, Liyuan Gao[2,] *, Victor Sheng[2], Jun Zhang[1]

Department of Molecular & Translational Medicine (MTM)[1]
Texas Tech University Health Science Center El Paso (TTUHSCEP), El Paso, TX 79905 USA
Department of Computer Sciences[2], Texas Tech University, Lubbock, TX 79409 USA

*Shared first authorship

Supplemental Materials

All correspondence:
Jun Zhang, Sc.D., Ph.D.
Department of Biomedical Sciences
Texas Tech University Health Science Center
5001 El Paso Drive, El Paso, TX 79905
Tel: (915) 215-4197 Email: jun.zhang2000@gmail.com

Suppl. Figure 1: **Multi-model machine learning utilized to predict functionality of protein isoforms**. **A.** Convolution Neural Network (CNN) with a threshold of 0.5 found a possible transcription factor in 60% of the testing repetitions. This result was consistent as it was the same sequence repeated in each of the positive repetitions **B.** CNN with a lower threshold: The unequal balance of transcription factors to non-transcription factors led the research team to lower the threshold to 0.1 to increase the sensitivity to the unbalanced nature of transcription factors to non-transcription factors. This resulted in more potential transcription factors being discovered in all repetitions of testing. **C.** A second model of machine learning was utilized in the form of Biased-SVM models. The research team repeated the process of starting with a threshold of 0.5 to predict any potential transcription factors and found multiple (n=8) in one testing repetition. **D.** Following the same procedure as the CNN the threshold for the biased-svm model had the threshold reduced to 0.1 to induce in-balance that is found in the protein-transcription ratio found in nature.

Suppl. Figure 2: **Multiple machine learning models found concurrent and unique proteins predicted to be transcription factors in multiple repetitions.** A Venn Diagram was utilized to compare the sequences found between the two models. The CNN and biased SVM models predicted 7 concurrent sequences to function as transcription factors, while each predicted their own unique sequences to also function as transcription factors.

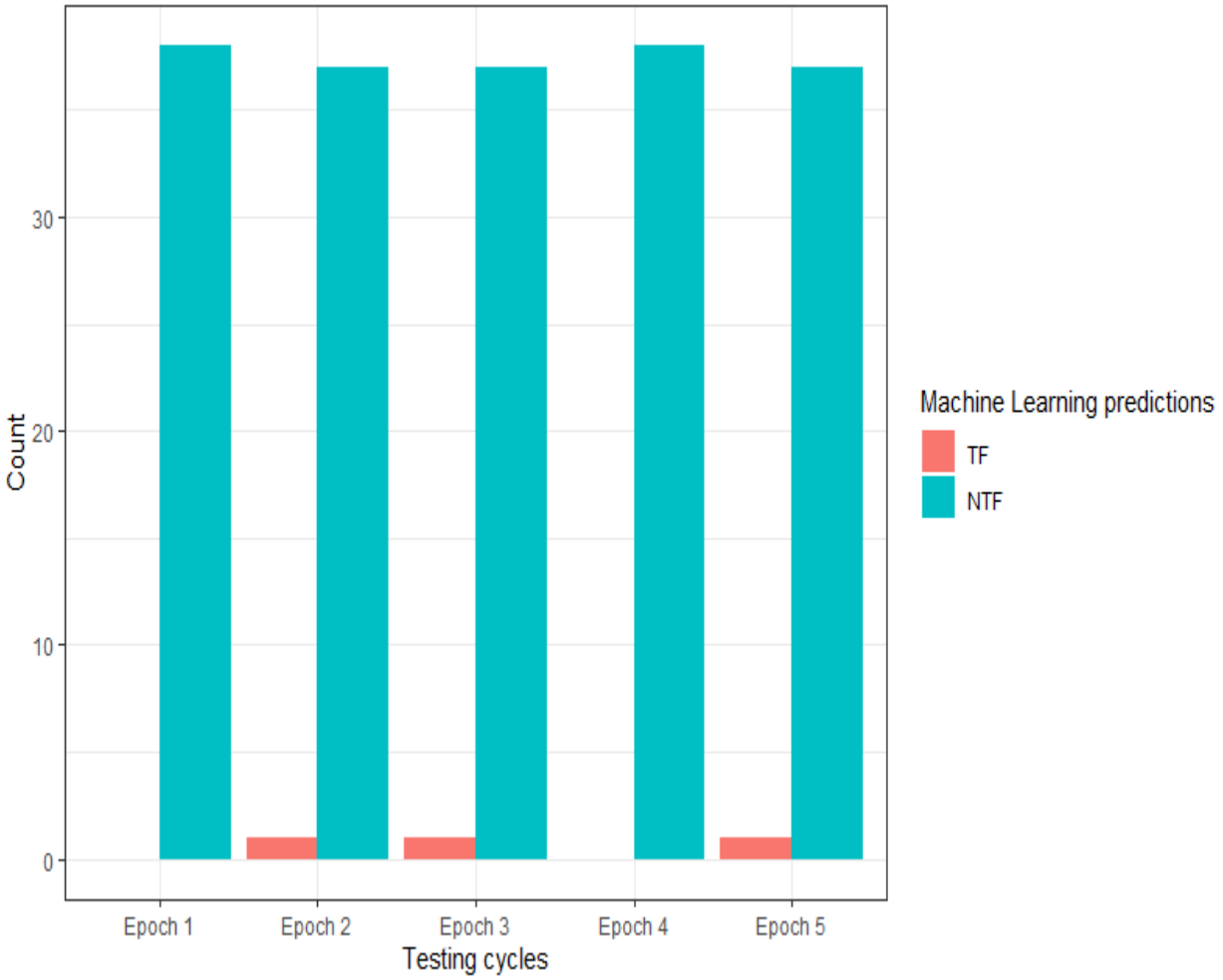
Suppl. Table 1: 5-Fold Cross Validations results of the CNN model used to test the sequence on a unique training model of 4331 unique FASTA sequences. The research team set a standard for all validations results to be at 95% to minimize all type 1 and type 2 errors. While the CNN did not deliver the F1 score of 95% we compared the results to the Biased-SVM as a way to verify findings.

Suppl. Table 2: 5-Fold Cross Validations results of the biased-SVM model used to test the sequence on a unique training model of 4331 unique FASTA sequences. Where the CNN model was lacking the 95% in all categories the biased-SVM model was able to accomplish this. We used the CNN model as a secondary comparison to the primary results found by the biased-SVM model to prevent any type 1 or type 2 errors.

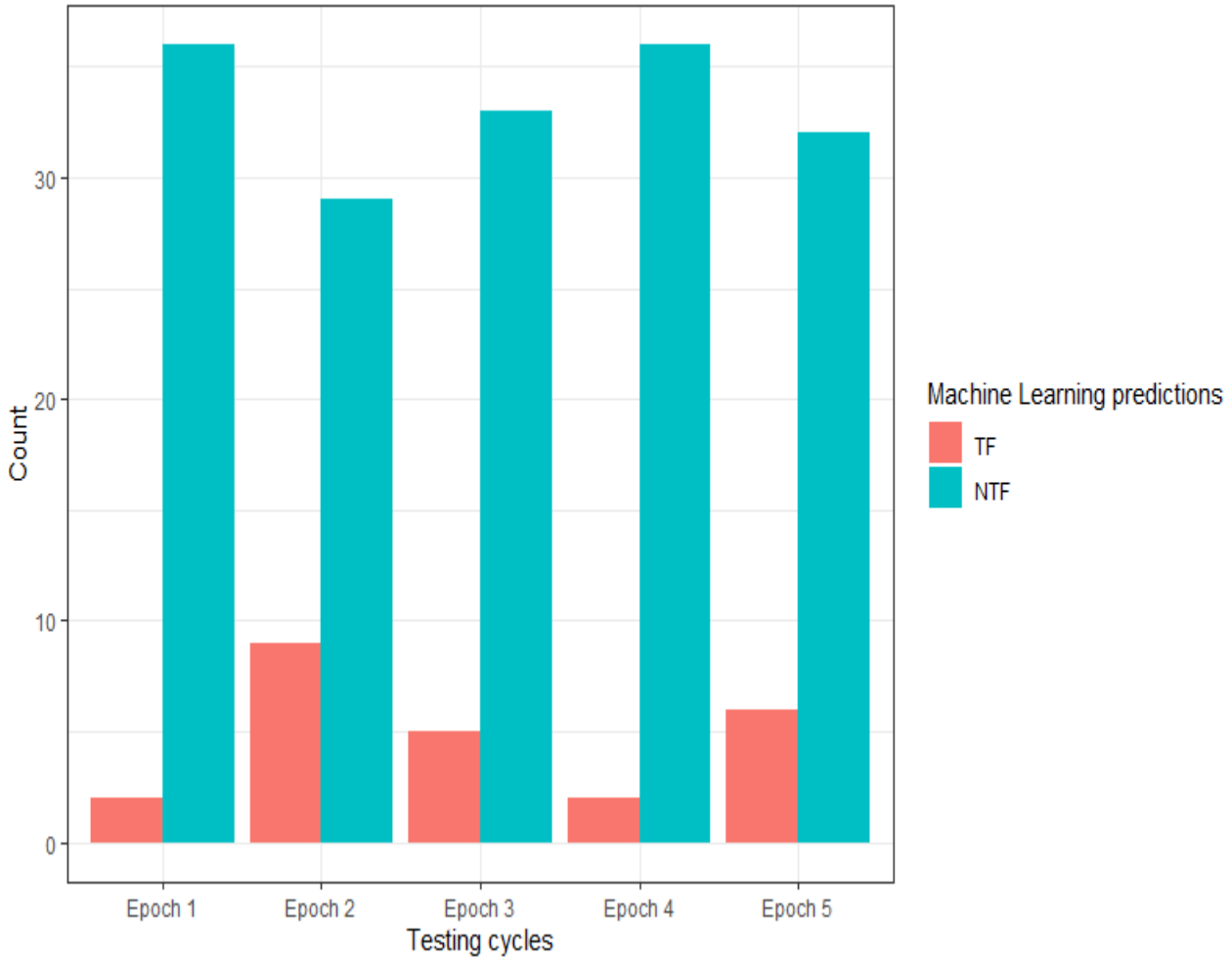
Other Suppl. Table containing all raw data for ML calculation, and will be provide by PI under the request.

Suppl. Figure 1:

1A: CNN Threshold ≥ 0.5

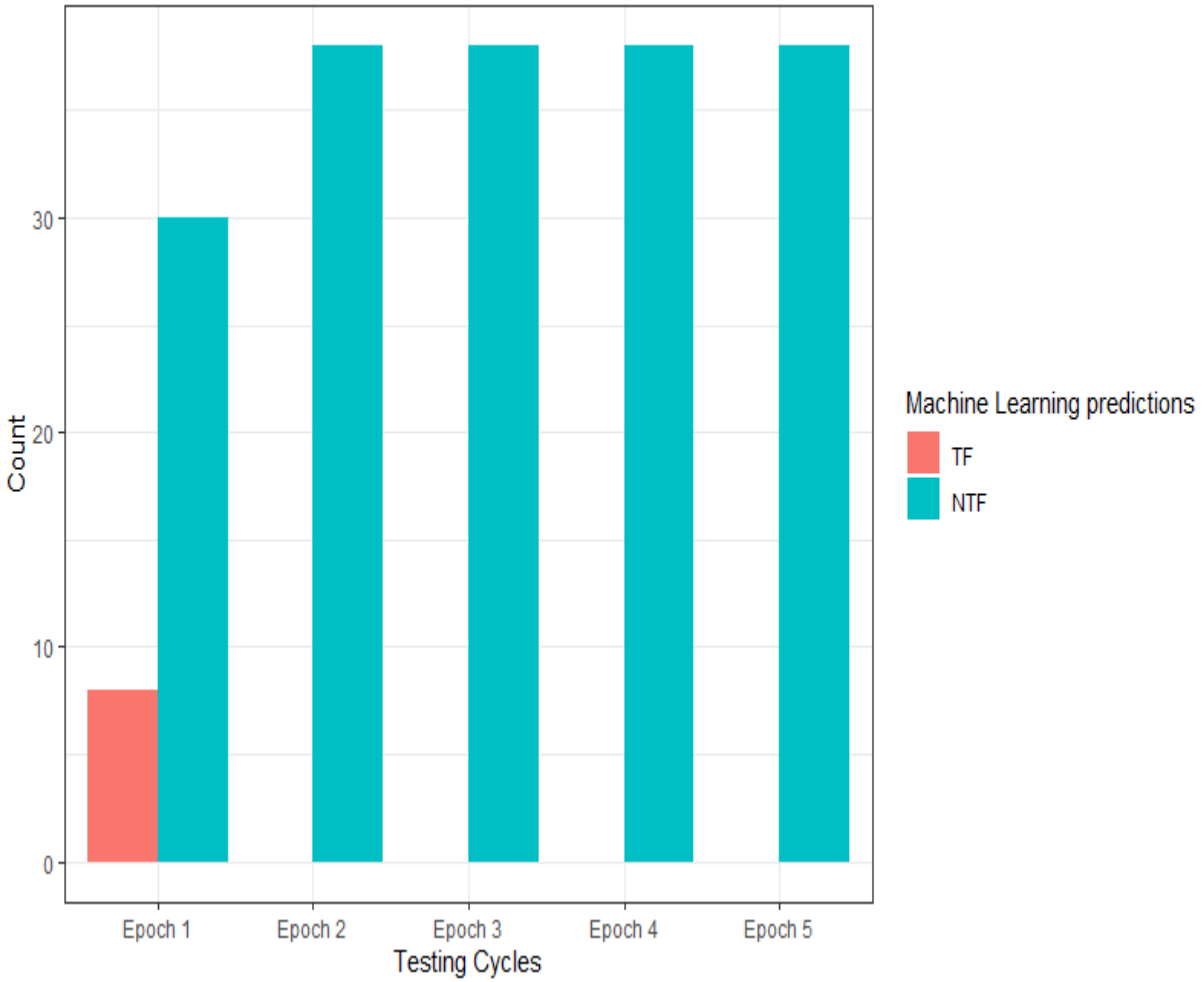


1B: CNN Threshold ≥ 0.1

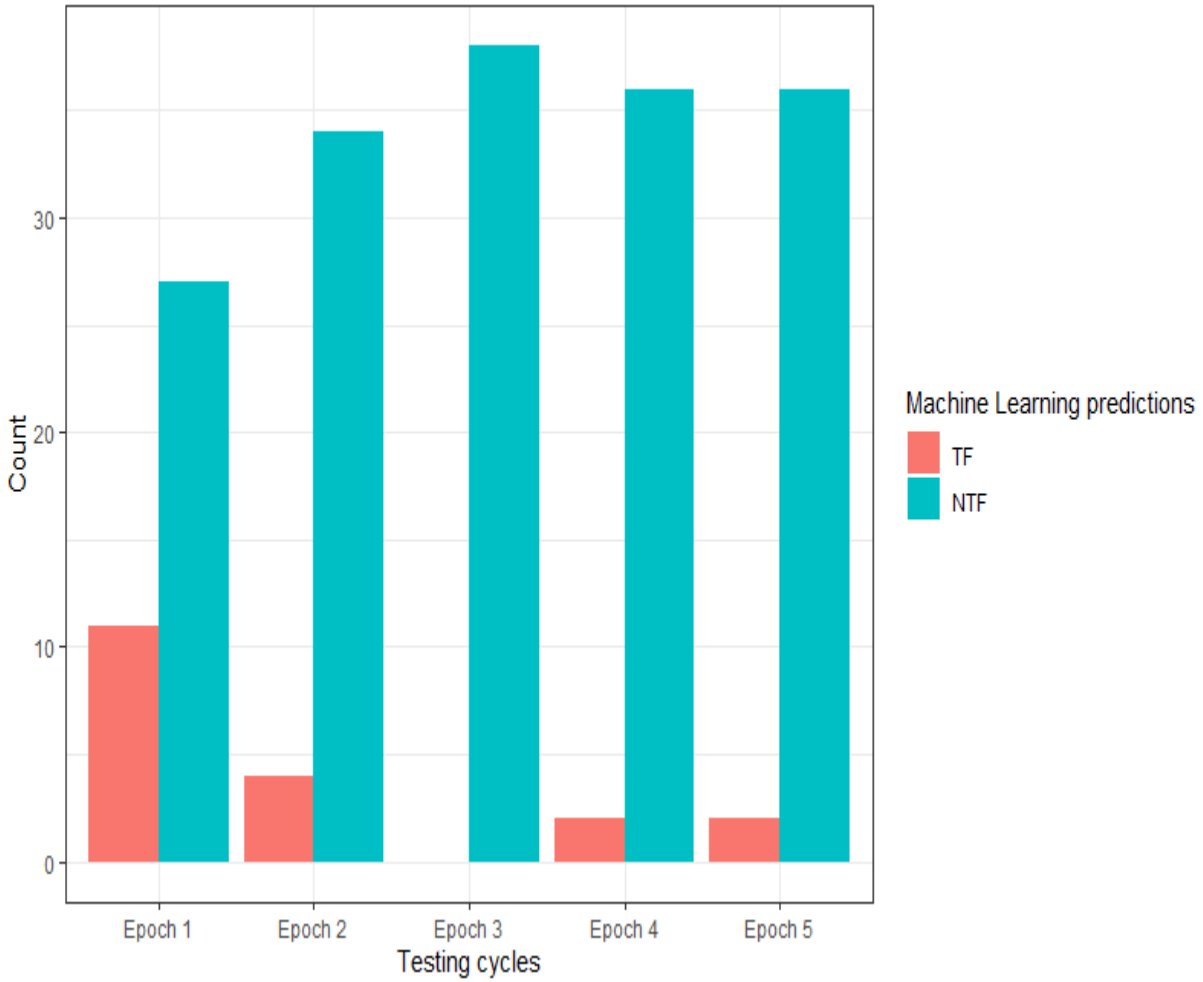


Suppl. Figure 1:

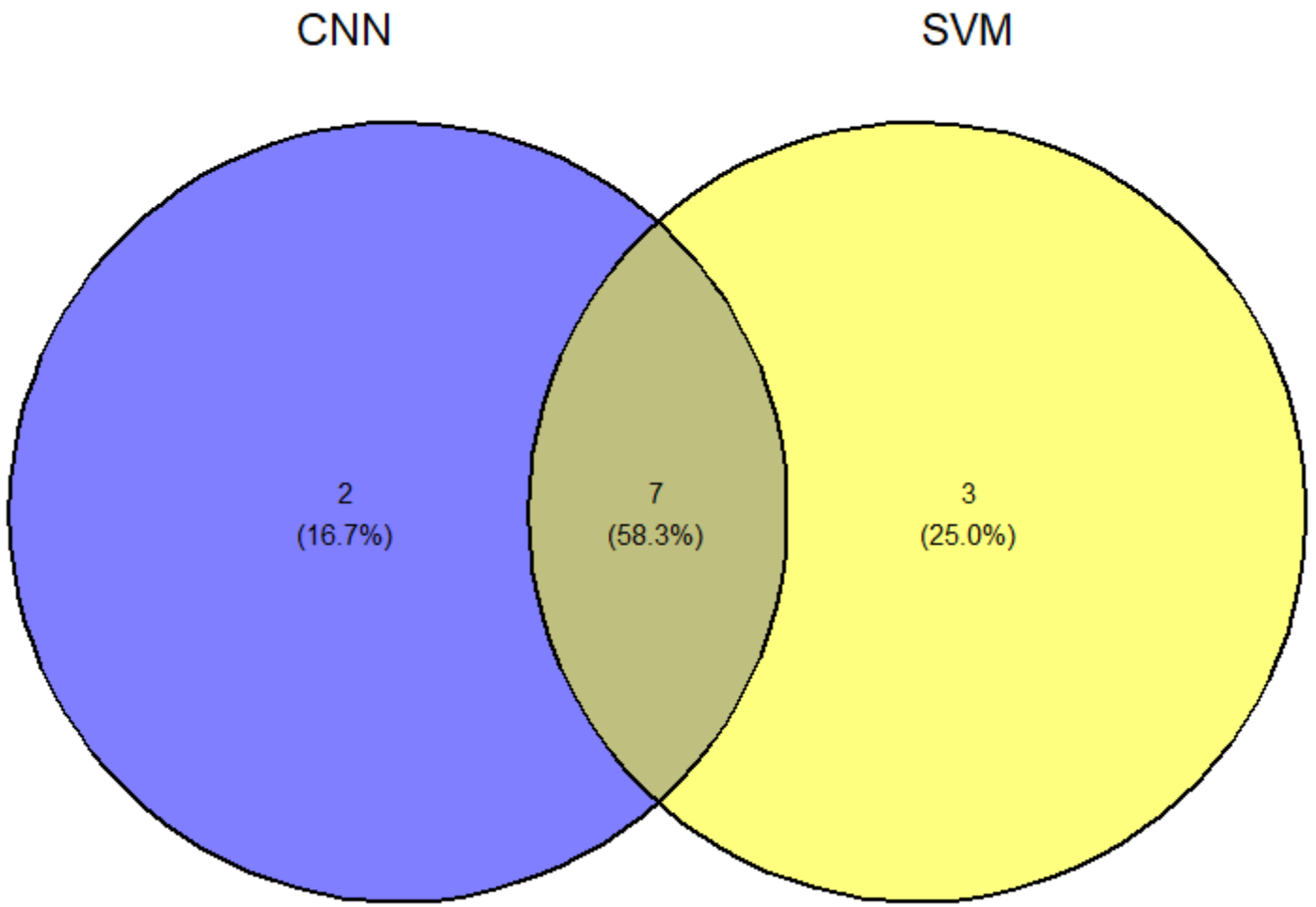
1C. SVM Threshold ≥ 0.5



1D: SVM Threshold ≥ 0.1



Suppl. Figure 2



Venn diagram of mutual and unique predictions per learning model technique

Suppl. Table 1: 5 Fold Cross Validation results of CNN model

CNN 5-Fold Cross validation results	
F1 Score	0.940
Specificity	0.956
Sensitivity	0.954
Balanced Accuracy	0.955

Suppl. Table 2: 5 Fold Cross Validation results of Biased-SVM model

Biased-SVM 5-Fold Cross Validation	
F1 Score	0.950
Specificity	0.962
Sensitivity	0.966
Balanced Accuracy	0.964