# DeepRMethylSite: Prediction of Arginine Methylation in Proteins using Deep Learning

Meenal Chaudhari[a+], Niraj Thapa[a+], Kaushik Roy[b], Robert H. Newman[c], Hiroto Saigo[d], Dukka B. KC[e*]

## Supplementary Information

Table S1: Comparison of CNN, LSTM and ensemble models (i.e., DeepRMethylSite) on the independent dataset used during the evaluation of PRmePred.

| Model | MCC | SN | SP | ACC | AUC |
|---|---|---|---|---|---|
| LSTM | 0.76 | 0.90 | 0.85 | 0.88 | 0.88 |
| CNN | 0.78 | 0.93 | 0.83 | 0.89 | 0.89 |
| DeepRMethylSite | 0.79 | 0.93 | 0.85 | 0.90 | 0.94 |

DeepRMethylSite used the optimized weights through the grid search method,[0.25,0.75] for LSTM and CNN models, respectively.

### XGBoost Performance on Feature Set

To construct a feature-based methylation site predictor, we extracted sequence-based features, such as Pseudo amino acid composition (PseAAC), Composition, Transition and Distribution (CTD), and Sequence Order Coupling Number (SOCN), from the dataset created using the FEPS server[24]. The best 500 features were selected using XGBOOST and used to develop feature-based models based on various popular machine learning algorithms. The resulting models were evaluated using our independent test for comparison with our ensemble deep learning model, DeepRMethylSite (Table S2).

Table S2: Performance metrics for various feature-based classifiers and our deep learning-based model, DeepRMethylSite, using the Independent Test. RF: Random Forest; SVM: Support Vector Machine; MCC: Matthew's Correlation Coefficient; SN: sensitivity; SP: Specificity; ACC: Accuracy.

| Model | MCC | SN | SP | ACC |
|---|---|---|---|---|
| RF | 0.43 | 0.63 | 0.79 | 0.71 |
| Naïve Bayes | 0.35 | 0.58 | 0.77 | 0.67 |
| XGBoost | 0.43 | 0.65 | 0.77 | 0.72 |
| SVM | 0.45 | **0.69** | 0.76 | 0.72 |
| DeepRMethylSite | **0.51** | 0.68 | **0.82** | **0.75** |

### Statistical analysis of deep learning-based predictors

In addition to 10-fold cross-validation, we also examined the differences in performance between the ensemble method and each of the component deep learning methods (i.e., the CNN- and LSTM-based methods) using the Student's t-test. To this end, we first divided our independent test into 10 equal, non-overlapping parts, such that the assumption of independence required for a Student's t-test was satisfied. The 10 subsets were then used to calculate the independent test results (Table S3). A paired Student's t-test was used to determine statistical significance between the methods. These analyses suggest that there is a statistically significant difference in MCC scores between LSTM and the ensemble method (*p*=0.002). Moreover, though not statistically significant, the difference in MCC between the CNN and the ensemble models appears to be trending toward significance (*p* = 0.20) (Table S3). In particular, there was a trend toward significance between CNN and the ensemble method with respect to specificity (*p* = 0.097).

Table S3: Independent Test Results using either CNN, LSTM or the ensemble model. MCC: Matthew's Correlation Coefficient; SN: sensitivity; SP: Specificity; OHE: One hot encoding; Emb: Embedding; ACC: Accuracy. Avg: Average; SE: Standard Error. *p*-values were calculated for the CNN or LSTM models versus the ensemble method using a paired Student's t-test.

| Model | CNN | | | LSTM | | | Ensemble | | |
|---|---|---|---|---|---|---|---|---|---|
| Subset | MCC | SP | SN | MCC | SP | SN | MCC | SP | SN |
| 1 | 0.420 | 0.760 | 0.654 | 0.400 | 0.750 | 0.649 | 0.431 | 0.769 | 0.659 |
| 2 | 0.510 | 0.779 | 0.736 | 0.510 | 0.813 | 0.692 | 0.564 | 0.817 | 0.745 |
| 3 | 0.490 | 0.774 | 0.716 | 0.440 | 0.769 | 0.668 | 0.501 | 0.779 | 0.721 |
| 4 | 0.440 | 0.808 | 0.630 | 0.370 | 0.793 | 0.563 | 0.446 | 0.817 | 0.620 |
| 5 | 0.440 | 0.764 | 0.673 | 0.430 | 0.769 | 0.654 | 0.424 | 0.750 | 0.673 |
| 6 | 0.490 | 0.813 | 0.678 | 0.490 | 0.856 | 0.620 | 0.517 | 0.837 | 0.673 |
| 7 | 0.530 | 0.832 | 0.692 | 0.500 | 0.827 | 0.663 | 0.520 | 0.832 | 0.683 |
| 8 | 0.540 | 0.832 | 0.702 | 0.520 | 0.856 | 0.659 | 0.530 | 0.837 | 0.688 |
| 9 | 0.490 | 0.764 | 0.726 | 0.410 | 0.740 | 0.668 | 0.501 | 0.779 | 0.721 |
| 10 | 0.490 | 0.831 | 0.653 | 0.460 | 0.798 | 0.662 | 0.495 | 0.826 | 0.662 |
| **Avg** | **0.484** | **0.796** | **0.686** | **0.453** | **0.797** | **0.650** | **0.493** | **0.804** | **0.684** |
| **SE** | **0.012** | **0.010** | **0.011** | **0.016** | **0.013** | **0.011** | **0.014** | **0.010** | **0.012** |
| ***p*-val** | **0.200** | **0.097** | **0.597** | **0.002** | **0.309** | **0.001** | -- | -- | -- |

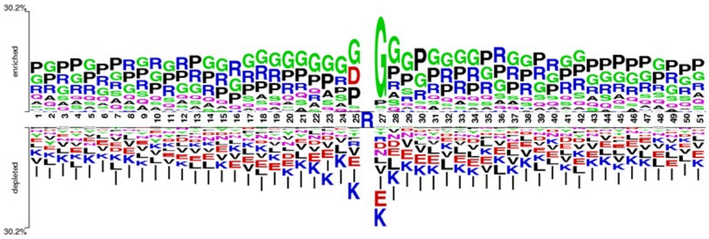
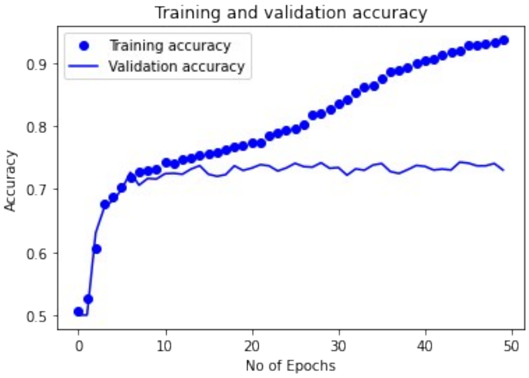


Figure S1: Two Logo Dataset on the Generated Dataset.

A. CNN



B. LSTM

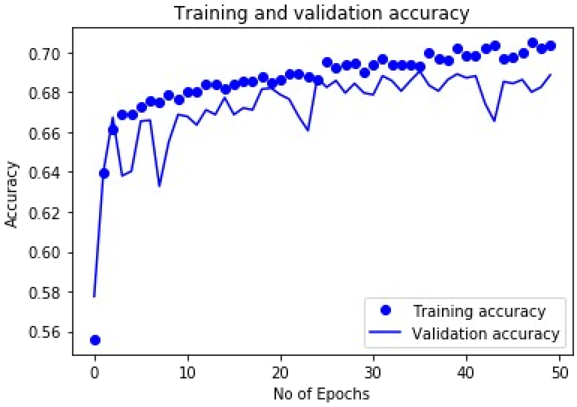


Fig S2: A. Plot showing accuracy vs number of epochs for CNN B. Plot showing accuracy vs number of epochs for LSTM models.