MBE, 19(12): 13294–13305.
DOI: 10.3934/mbe.2022622
Received: 04 July 2022

Revised: 23 August 2022
Accepted: 29 August 2022
Published: 13 September 2022

http://www.aimspress.com/journal/MBE

***Research article***

**SPREAD: An ensemble predictor based on DNA autoencoder framework for discriminating promoters in Pseudomonas aeruginosa**

**Shengming Zhou, Jia Zheng\* and Cangzhi Jia\***

School of Science, Dalian Maritime University, Dalian 116026, China

**\*Correspondence:** Email: zhengjia@dlmu.edu.cn, cangzhijia@dlmu.edu.cn.

**Supplementary**

**Table S1.** Parameter setting of eXtreme Gradient Boosting model.

|  |  |
| --- | --- |
| Parameter | Parameter setting |
| learning\_rate | 0.3 |
| min\_child\_weight | 1 |
| max\_depth | 6 |
| gamma | 0 |
| subsample | 1 |
| colsample\_bytree | 1 |
| reg\_lambda | 1 |
| n\_estimators | 300 |
| random\_state | 42 |

**Table S2.** Parameter setting of random forest model.

|  |  |
| --- | --- |
| Parameter | Parameter setting |
| n\_estimators | 200 |
| random\_state | 42 |

**Table S3.** Parameter setting of support vector machine model.

|  |  |
| --- | --- |
| Parameter | Parameter setting |
| kernel | ‘rbf’ |
| C | 1 |
| gamma | 1 |
| probability | True |

**Table S4.** Parameter setting of k nearest neighbors mode.

|  |  |
| --- | --- |
| Parameter | Parameter setting |
| n\_neighbors | 10 |

**Table S5.** Parameter setting of CNN model.

|  |  |
| --- | --- |
| Network | Parameter setting |
| Conv1D\_1 | Filters = 128, conv\_size = 7, padding = ‘same’, activation = ‘relu’ |
| MaxPooling1D\_1 | max\_pooling\_size = 2 |
| Dropout\_1 | 0.5 |
| Conv1D\_2 | Filters = 128, conv\_size = 7, padding = ‘same’, activation = ‘relu’ |
| MaxPooling1D\_2 | max\_pooling\_size = 2 |
| Dropout\_2 | 0.5 |
| Flatten |  |
| Activation | ‘relu’ |
| Dense | latent\_dim = 256 |

**Table S6.** The reslut of the integration of two kinds of models by calculating the mean of the output probability.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Integration strategy | Recall | Precision | F1-score | Accuracy | AUC |
| CNN+XGBoost | 0.5 | 0.92(±0.02) | 0.85(±0.02) | 0.88(±0.01) | 0.88(±0.01) | 0.95(±0.01) |
| **CNN+RF** | **0.5** | **0.91(±0.03)** | **0.88(±0.02)** | **0.89(±0.02)** | **0.89(±0.02)** | **0.95(±0.01)** |
| CNN+GNB | 0.5 | 0.76(±0.03) | 0.82(±0.02) | 0.79(±0.02) | 0.80(±0.02) | 0.93(±0.01) |
| CNN+SVM | 0.5 | 0.91(±0.02) | 0.87(±0.02) | 0.89(±0.02) | 0.89(±0.02) | 0.95(±0.01) |
| CNN+KNN | 0.5 | 0.91(±0.02) | 0.87(±0.02) | 0.89(±0.02) | 0.89(±0.02) | 0.95(±0.01) |
| CNN+RF+XGBoost | 0.5 | 0.90(±0.02) | 0.85(±0.02) | 0.87(±0.01) | 0.87(±0.01) | 0.94(±0.01) |
| CNN+RF+GNB | 0.5 | 0.84(±0.03) | 0.85(±0.02) | 0.85(±0.02) | 0.85(±0.02) | 0.93(±0.01) |
| CNN+RF+SVM | 0.5 | 0.91(±0.02) | 0.86(±0.02) | 0.88(±0.01) | 0.88(±0.02) | 0.94(±0.01) |
| CNN+RF+KNN | 0.5 | 0.90(±0.03) | 0.86(±0.02) | 0.88(±0.02) | 0.88(±0.02) | 0.94(±0.01) |
|  |  |  |  |  | *Continued on next page* |
| Integration strategy |  | Recall | Precision | F1-score | Accuracy | AUC |
| CNN+XGBoost | 0.55 | 0.88(±0.02) | 0.87(±0.02) | 0.88(±0.02) | 0.87(±0.02) | 0.95(±0.01) |
| CNN+RF | 0.55 | 0.88(±0.03) | 0.89(±0.02) | 0.89(±0.02) | 0.89(±0.02) | 0.95(±0.01) |
| CNN+GNB | 0.55 | 0.73(±0.03) | 0.85(±0.02) | 0.78(±0.02) | 0.80(±0.02) | 0.93(±0.01) |
| CNN+SVM | 0.55 | 0.89(±0.03) | 0.89(±0.02) | 0.89(±0.02) | 0.89(±0.02) | 0.95(±0.01) |
| CNN+KNN | 0.55 | 0.89(±0.03) | 0.88(±0.02) | 0.89(±0.01) | 0.89(±0.01) | 0.95(±0.01) |
| CNN+KNN+XGBoost | 0.55 | 0.88(±0.03) | 0.85(±0.02) | 0.87(±0.02) | 0.87(±0.02) | 0.94(±0.01) |
| CNN+KNN+RF | 0.55 | 0.88(±0.02) | 0.87(±0.02) | 0.88(±0.02) | 0.88(±0.02) | 0.94(±0.01) |
| CNN+KNN+GNB | 0.55 | 0.81(±0.03) | 0.85(±0.02) | 0.83(±0.02) | 0.83(±0.02) | 0.93(±0.01) |
| CNN+KNN+SVM | 0.55 | 0.89(±0.03) | 0.87(±0.02) | 0.88(±0.02) | 0.87(±0.02) | 0.94(±0.01) |
| CNN+XGBoost | 0.6 | 0.85(±0.03) | 0.88(±0.02) | 0.87(±0.02) | 0.87(±0.02) | 0.95(±0.01) |
| CNN+RF | 0.6 | 0.84(±0.03) | 0.91(±0.02) | 0.87(±0.01) | 0.88(±0.01) | 0.95(±0.01) |
| CNN+GNB | 0.6 | 0.72(±0.04) | 0.86(±0.03) | 0.78(±0.02) | 0.80(±0.02) | 0.93(±0.01) |
| CNN+SVM | 0.6 | 0.86(±0.04) | 0.90(±0.02) | 0.88(±0.02) | 0.88(±0.02) | 0.95(±0.01) |
| CNN+KNN | 0.6 | 0.86(±0.03) | 0.90(±0.02) | 0.88(±0.02) | 0.88(±0.02) | 0.95(±0.01) |
| CNN+RF+XGBoost | 0.6 | 0.86(±0.03) | 0.87(±0.02) | 0.86(±0.02) | 0.86(±0.02) | 0.94(±0.01) |
| CNN+RF+GNB | 0.6 | 0.76(±0.03) | 0.86(±0.02) | 0.81(±0.02) | 0.82(±0.01) | 0.93(±0.01) |
| CNN+RF+SVM | 0.6 | 0.85(±0.03) | 0.88(±0.02) | 0.86(±0.02) | 0.87(±0.02) | 0.94(±0.01) |
| CNN+RF+KNN | 0.6 | 0.85(±0.03) | 0.88(±0.02) | 0.86(±0.02) | 0.87(±0.02) | 0.94(±0.01) |
| CNN+XGBoost | stacking | 0.89(±0.03) | 0.83(±0.02) | 0.86(±0.02) | 0.86(±0.02) | 0.86(±0.02) |
| CNN+RF | stacking | 0.91(±0.03) | 0.87(±0.02) | 0.89(±0.02) | 0.89(±0.02) | 0.89(±0.02) |
| CNN+GNB | stacking | 0.88(±0.02) | 0.88(±0.02) | 0.88(±0.01) | 0.88(±0.01) | 0.88(±0.01) |
| CNN+SVM | stacking | 0.89(±0.02) | 0.88(±0.02) | 0.88(±0.01) | 0.88(±0.01) | 0.88(±0.01) |
| CNN+KNN | stacking | 0.89(±0.02) | 0.88(±0.02) | 0.88(±0.01) | 0.88(±0.01) | 0.88(±0.01) |
| CNN+RF+XGBoost | stacking | 0.89(±0.02) | 0.83(±0.02) | 0.86(±0.01) | 0.86(±0.01) | 0.86(±0.01) |
| CNN+RF+GNB | stacking | 0.92(±0.02) | 0.86(±0.02) | 0.89(±0.02) | 0.89(±0.02) | 0.89(±0.02) |
| CNN+RF+SVM | stacking | 0.92(±0.02) | 0.87(±0.02) | 0.89(±0.02) | 0.89(±0.02) | 0.89(±0.02) |
| CNN+RF+KNN | stacking | 0.92(±0.03) | 0.87(±0.02) | 0.89(±0.02) | 0.89(±0.02) | 0.89(±0.02) |
| CNN+RF+SVM+XGBoost | stacking | 0.89(±0.02) | 0.83(±0.02) | 0.86(±0.01) | 0.86(±0.01) | 0.86(±0.01) |
| CNN+RF+SVM+GNB | stacking | 0.92(±0.03) | 0.86(±0.02) | 0.89(±0.02) | 0.89(±0.02) | 0.89(±0.02) |
| CNN+RF+SVM+KNN | stacking | 0.92(±0.02) | 0.87(±0.02) | 0.89(±0.02) | 0.89(±0.02) | 0.89(±0.02) |
| CNN+RF+SVM+KNN+XGBoost | stacking | 0.89(±0.02) | 0.83(±0.02) | 0.86(±0.01) | 0.86(±0.01) | 0.86(±0.01) |
| CNN+RF+SVM+KNN+GNB | stacking | 0.92(±0.03) | 0.86(±0.02) | 0.89(±0.02) | 0.89(±0.02) | 0.89(±0.02) |

**Table S7.** Compare the experimental results.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | Best size | Recall | Precision | Specificity | F1-score | Accuracy |
| G4PromFinder | - | 0.69 | 0.431 | 0.089 | 0.53 | 0.389 |
| CNN+RF | 81\*32, 256 | 0.98(±0.00) | 0.98(±0.02) | 0.97(±0.02) | 0.98(±0.01) | 0.98(±0.01) |
| CNN | 81\*32 | 0.95(±0.01) | 0.98(±0.01) | 0.98(±0.01) | 0.96(±0.01) | 0.97(±0.01) |
| XGBoost | 512 | 0.99(±0.00) | 0.97(±0.03) | 0.97(±0.03) | 0.98(±0.02) | 0.98(±0.02) |
| RF | 256 | 0.99(±0.00) | 0.97(±0.03) | 0.97(±0.03) | 0.98(±0.01) | 0.98(±0.02) |
| GNB | 512 | 0.73(±0.02) | 0.80(±0.02) | 0.82(±0.02) | 0.76(±0.02) | 0.77(±0.01) |
| SVM | 512 | 0.88(±0.01) | 0.85(±0.01) | 0.84(±0.02) | 0.86(±0.01) | 0.86(±0.01) |
| KNN | 512 | 0.90(±0.01) | 0.84(±0.02) | 0.83(±0.02) | 0.87(±0.01) | 0.87(±0.01) |

**Table S8.** The results of six individual classification algorithms on the training dataset.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Recall | Precision | F1\_score | Accuracy | AUC |
| CNN | 0.88(±0.06) | 0.81(±0.06) | 0.84(±0.02) | 0.83(±0.03) | 0.93(±0.01) |
| XGBoost | 0.80(±0.02) | 0.79(±0.02) | 0.79(±0.02) | 0.79(±0.02) | 0.85(±0.01) |
| RF | 0.79(±0.02) | 0.80(±0.02) | 0.80(±0.02) | 0.80(±0.02) | 0.85(±0.02) |
| GNB | 0.74(±0.02) | 0.72(±0.02) | 0.73(±0.02) | 0.72(±0.02) | 0.78(±0.02) |
| SVM | 0.83(±0.02) | 0.80(±0.02) | 0.81(±0.01) | 0.81(±0.01) | 0.87(±0.02) |
| KNN | 0.84(±0.02) | 0.79(±0.02) | 0.81(±0.01) | 0.81(±0.02) | 0.85(±0.01) |

© 2022 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0)