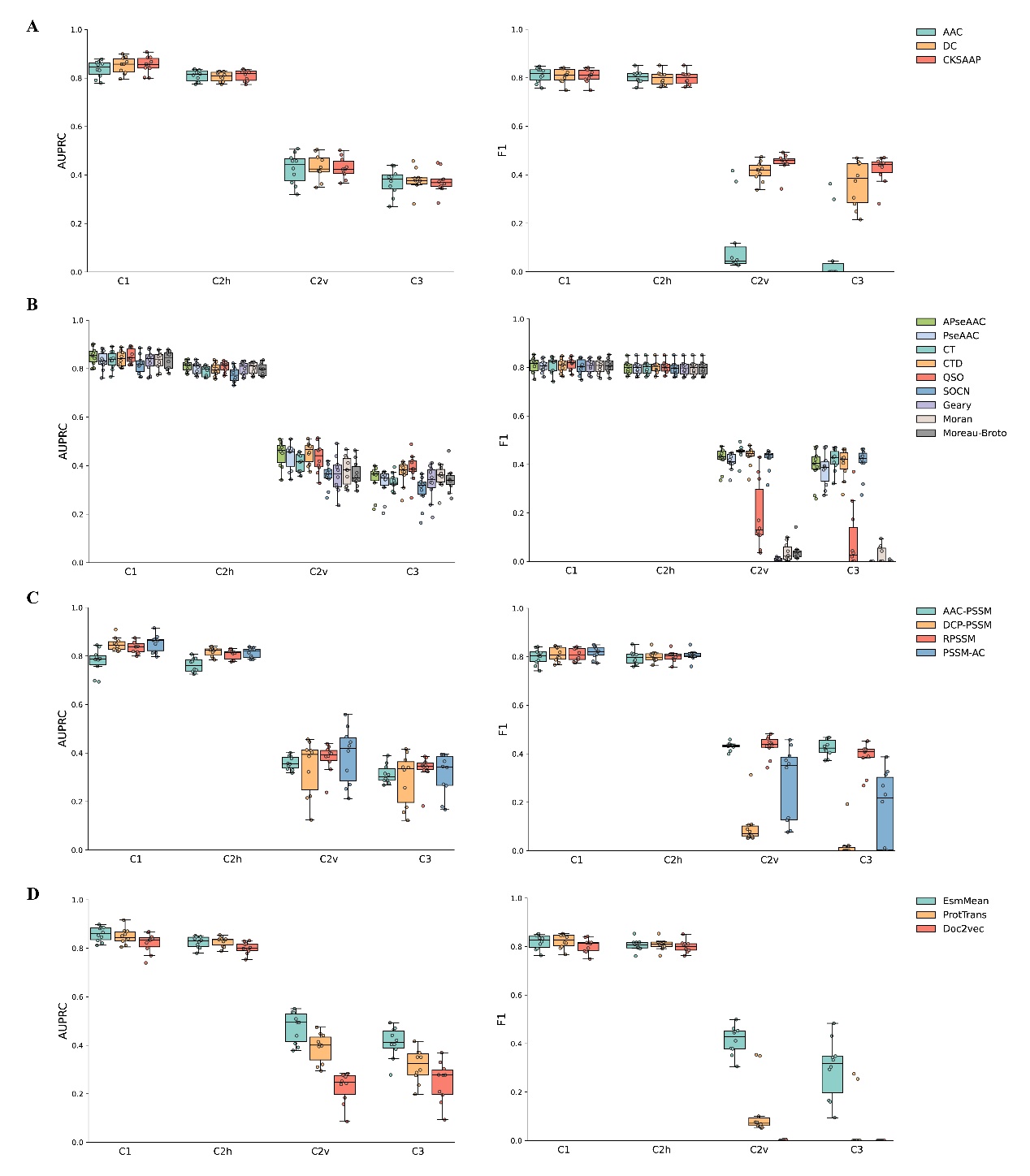
**Supplementary Figure 1**

**Supplementary Fig. 1. AUPRC and F1 of RF-based encoding schemes for different groups in C1, C2h, C2v and C3 test sets.**

**Supplementary Table 3. The details of the C1, C2v, C2h and C3 data sets**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | **Train set** | **C1 test set** | **C2v test set** | **C2h test set** | **C3 test set** |
| 1 | total | 9844 | 1094 | 4795 | 4706 | 2122 |
|  | pos | 918 | 102 | 427 | 406 | 198 |
|  | neg | 8926 | 992 | 4368 | 4300 | 1924 |
| 2 | total | 9824 | 1092 | 4793 | 4761 | 2091 |
|  | pos | 856 | 95 | 390 | 498 | 212 |
|  | neg | 8968 | 997 | 4403 | 4263 | 1879 |
| 3 | total | 9782 | 1087 | 4780 | 4825 | 2087 |
|  | pos | 910 | 101 | 436 | 423 | 181 |
|  | neg | 8872 | 986 | 4344 | 4402 | 1906 |
| 4 | total | 10237 | 1138 | 4687 | 4600 | 1899 |
|  | pos | 940 | 105 | 444 | 400 | 162 |
|  | neg | 9297 | 1033 | 4243 | 4200 | 1737 |
| 5 | total | 9643 | 1072 | 4738 | 5037 | 2071 |
|  | pos | 909 | 101 | 429 | 434 | 178 |
|  | neg | 8734 | 971 | 4309 | 4603 | 1893 |
| 6 | total | 10013 | 1113 | 4659 | 4761 | 2015 |
|  | pos | 929 | 103 | 428 | 418 | 173 |
|  | neg | 9084 | 1010 | 4231 | 4343 | 1842 |
| 7 | total | 10267 | 1141 | 4834 | 4442 | 1877 |
|  | pos | 884 | 98 | 427 | 460 | 182 |
|  | neg | 9383 | 1043 | 4407 | 3982 | 1695 |
| 8 | total | 10207 | 1135 | 4725 | 4602 | 1892 |
|  | pos | 942 | 105 | 431 | 397 | 176 |
|  | neg | 9265 | 1030 | 4294 | 4205 | 1716 |
| 9 | total | 9720 | 1080 | 4696 | 4982 | 2083 |
|  | pos | 946 | 105 | 425 | 400 | 175 |
|  | neg | 8774 | 975 | 4271 | 4582 | 1908 |
| 10 | total | 10257 | 1140 | 4546 | 4644 | 1974 |
|  | pos | 899 | 100 | 428 | 434 | 190 |
|  | neg | 9358 | 1040 | 4118 | 4210 | 1784 |

**Supplementary Table 4. Parameter setting and optimization of different models based on the ML method**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Parameters** | **Parameter search space** | **Package** |
| RF | n\_estimators=1000 max\_depth=18 max\_features='sqrt' criterion='entropy' bootstrap=True | [10, 50, 100, 200, 500, 700, 800, 1000] [2, 4, 6, 8, 10, 12, 14, 15, 16, 18, 20] ['sqrt', 'log2'] ['gini', 'entropy'] [True] | scikit-learn |
| XGB | n\_estimators=100 learning\_rate=0.05 max\_depth=15 min\_child\_weight=1 gamma=0.3 colsample\_bytree=0.3 reg\_alpha=0 reg\_lambda=1 | [100, 200, 300, 400, 500] [0.01, 0.05, 0.1, 0.15, 0.2] [5, 6, 8, 10, 12, 15] [1, 3, 5, 7] [0.0, 0.1, 0.2, 0.3, 0.4] [0.3, 0.4, 0.5, 0.7, 0.9] [0, 0.1, 0.5, 1] [0, 0.1, 0.5, 1] | xgboost |
| SVM | C=10 gamma=0.01 kernel='rbf' class\_weight='balanced' | [0.1, 1, 10, 100] [0.1, 1, 5, 10, 50, 100] ['linear', 'rbf', 'poly'] ['balanced'] | scikit-learn |
| AB | n\_estimators=50 learning\_rate=0.1 algorithm='SAMME' | [10, 50, 100, 200] [0.01, 0.1, 1] ['SAMME', 'SAMME.R'] | scikit-learn |
| MLP | optimizers=Adam learning\_rate=0.001 epochs=15 batch\_size=16 | [Adam] [0.001, 0.01, 0.1] [10, 15, 20, 30, 50, 80] [8, 16, 32, 64, 128] | tensorflow |

**Supplementary Table 6. Performance of RF-based encoding schemes for different groups in C1, C2h, C2v and C3 test sets**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Groups** | **Test set** | **Feature** | **Precision** | **Recall** | **Specificity** | **F1** | **AUROC** | **AUPRC** |
| Amino acid composition and order | C1 | AAC | 0.853±0.030 | 0.774±0.040 | 0.986±0.003 | 0.811±0.029 | 0.974±0.008 | 0.837±0.032 |
|  |  | DC | 0.854±0.026 | 0.769±0.042 | 0.987±0.003 | 0.809±0.028 | 0.978±0.008 | 0.852±0.032 |
|  |  | CKSAAP | 0.855±0.027 | 0.768±0.042 | 0.987±0.003 | 0.808±0.027 | 0.979±0.006 | 0.854±0.033 |
|  | C2h | AAC | 0.836±0.017 | 0.773±0.037 | 0.985±0.002 | 0.803±0.024 | 0.959±0.008 | 0.809±0.022 |
|  |  | DC | 0.835±0.015 | 0.766±0.041 | 0.985±0.002 | 0.799±0.027 | 0.956±0.006 | 0.806±0.019 |
|  |  | CKSAAP | 0.836±0.015 | 0.766±0.041 | 0.985±0.001 | 0.799±0.026 | 0.957±0.006 | 0.811±0.021 |
|  | C2v | AAC | 0.797±0.181 | 0.078±0.107 | 0.995±0.008 | 0.118±0.141 | 0.767±0.056 | 0.426±0.060 |
|  |  | DC | 0.600±0.105 | 0.326±0.047 | 0.974±0.019 | 0.415±0.038 | 0.790±0.034 | 0.433±0.050 |
|  |  | CKSAAP | 0.561±0.081 | 0.380±0.026 | 0.968±0.016 | 0.450±0.039 | 0.775±0.035 | 0.429±0.041 |
|  | C3 | AAC | 0.210±0.342 | 0.051±0.098 | 0.995±0.010 | 0.071±0.132 | 0.725±0.066 | 0.371±0.053 |
|  |  | DC | 0.598±0.132 | 0.284±0.093 | 0.973±0.029 | 0.364±0.089 | 0.750±0.030 | 0.379±0.044 |
|  |  | CKSAAP | 0.511±0.091 | 0.364±0.035 | 0.961±0.024 | 0.422±0.055 | 0.728±0.040 | 0.373±0.046 |
| Amino acid physicochemical properties | C1 | APAAC | 0.852±0.030 | 0.772±0.041 | 0.987±0.003 | 0.810±0.030 | 0.976±0.008 | 0.851±0.034 |
|  |  | PAAC | 0.850±0.023 | 0.766±0.039 | 0.986±0.002 | 0.806±0.024 | 0.974±0.008 | 0.833±0.036 |
|  |  | CT | 0.852±0.028 | 0.770±0.047 | 0.987±0.003 | 0.808±0.031 | 0.976±0.008 | 0.835±0.035 |
|  |  | CTD | 0.849±0.025 | 0.772±0.038 | 0.986±0.003 | 0.808±0.026 | 0.975±0.009 | 0.841±0.030 |
|  |  | QSO | 0.856±0.027 | 0.779±0.035 | 0.987±0.003 | 0.815±0.023 | 0.977±0.007 | 0.851±0.028 |
|  |  | SOCN | 0.849±0.025 | 0.762±0.045 | 0.986±0.002 | 0.803±0.030 | 0.965±0.011 | 0.817±0.037 |
|  |  | Geary | 0.854±0.027 | 0.763±0.042 | 0.987±0.003 | 0.805±0.027 | 0.975±0.007 | 0.832±0.041 |
|  |  | Moran | 0.853±0.027 | 0.761±0.044 | 0.987±0.003 | 0.804±0.028 | 0.975±0.007 | 0.833±0.036 |
|  |  | Moreau-Broto | 0.853±0.025 | 0.768±0.045 | 0.987±0.003 | 0.808±0.028 | 0.975±0.008 | 0.835±0.035 |
|  | C2h | APAAC | 0.833±0.017 | 0.769±0.039 | 0.985±0.002 | 0.799±0.025 | 0.957±0.006 | 0.810±0.020 |
|  |  | PAAC | 0.833±0.015 | 0.767±0.040 | 0.985±0.002 | 0.798±0.025 | 0.955±0.008 | 0.803±0.023 |
|  |  | CT | 0.836±0.014 | 0.766±0.041 | 0.985±0.001 | 0.799±0.026 | 0.950±0.006 | 0.790±0.021 |
|  |  | CTD | 0.832±0.017 | 0.773±0.037 | 0.984±0.002 | 0.801±0.023 | 0.952±0.007 | 0.797±0.025 |
|  |  | QSO | 0.834±0.015 | 0.772±0.037 | 0.985±0.002 | 0.802±0.023 | 0.959±0.006 | 0.808±0.020 |
|  |  | SOCN | 0.831±0.016 | 0.765±0.040 | 0.985±0.002 | 0.796±0.025 | 0.943±0.008 | 0.775±0.028 |
|  |  | Geary | 0.834±0.016 | 0.763±0.042 | 0.985±0.001 | 0.797±0.027 | 0.953±0.006 | 0.801±0.025 |
|  |  | Moran | 0.834±0.015 | 0.763±0.042 | 0.985±0.001 | 0.796±0.027 | 0.953±0.007 | 0.802±0.024 |
|  |  | Moreau-Broto | 0.833±0.017 | 0.765±0.042 | 0.985±0.002 | 0.797±0.027 | 0.952±0.007 | 0.797±0.024 |
|  | C2v | APAAC | 0.543±0.109 | 0.354±0.026 | 0.965±0.025 | 0.423±0.043 | 0.833±0.039 | 0.448±0.050 |
|  |  | PAAC | 0.550±0.108 | 0.340±0.037 | 0.966±0.026 | 0.412±0.033 | 0.829±0.042 | 0.438±0.051 |
|  |  | CT | 0.566±0.053 | 0.376±0.022 | 0.971±0.009 | 0.451±0.029 | 0.768±0.034 | 0.411±0.035 |
|  |  | CTD | 0.536±0.095 | 0.372±0.029 | 0.963±0.023 | 0.434±0.038 | 0.828±0.045 | 0.452±0.043 |
|  |  | QSO | 0.757±0.164 | 0.120±0.103 | 0.994±0.007 | 0.187±0.133 | 0.773±0.050 | 0.433±0.057 |
|  |  | SOCN | 0.481±0.090 | 0.384±0.029 | 0.953±0.028 | 0.422±0.041 | 0.734±0.039 | 0.358±0.045 |
|  |  | Geary | 0.600±0.357 | 0.004±0.003 | 1.000±0.000 | 0.008±0.006 | 0.688±0.063 | 0.363±0.071 |
|  |  | Moran | 0.786±0.223 | 0.019±0.018 | 1.000±0.000 | 0.037±0.034 | 0.686±0.055 | 0.380±0.056 |
|  |  | Moreau-Broto | 0.781±0.193 | 0.023±0.020 | 0.999±0.001 | 0.043±0.035 | 0.708±0.056 | 0.366±0.050 |
|  | C3 | APAAC | 0.490±0.132 | 0.336±0.042 | 0.956±0.034 | 0.390±0.069 | 0.744±0.048 | 0.342±0.059 |
|  |  | PAAC | 0.499±0.131 | 0.322±0.056 | 0.958±0.037 | 0.379±0.065 | 0.727±0.050 | 0.326±0.057 |
|  |  | CT | 0.507±0.063 | 0.359±0.039 | 0.964±0.011 | 0.419±0.043 | 0.686±0.032 | 0.334±0.031 |
|  |  | CTD | 0.478±0.111 | 0.360±0.037 | 0.953±0.031 | 0.404±0.058 | 0.754±0.046 | 0.369±0.048 |
|  |  | QSO | 0.466±0.419 | 0.060±0.089 | 0.995±0.009 | 0.089±0.124 | 0.740±0.059 | 0.390±0.055 |
|  |  | SOCN | 0.453±0.103 | 0.378±0.028 | 0.948±0.031 | 0.407±0.059 | 0.676±0.045 | 0.296±0.062 |
|  |  | Geary | 0.000±0.000 | 0.000±0.000 | 1.000±0.000 | 0.000±0.000 | 0.677±0.062 | 0.331±0.068 |
|  |  | Moran | 0.400±0.490 | 0.013±0.018 | 1.000±0.000 | 0.026±0.034 | 0.676±0.044 | 0.354±0.045 |
|  |  | Moreau-Broto | 0.100±0.300 | 0.000±0.001 | 1.000±0.000 | 0.001±0.003 | 0.695±0.056 | 0.342±0.050 |
| Evolutionary information | C1 | AAC-PSSM | 0.851±0.026 | 0.755±0.048 | 0.987±0.003 | 0.799±0.031 | 0.954±0.013 | 0.779±0.048 |
|  |  | DPC-PSSM | 0.846±0.027 | 0.779±0.038 | 0.986±0.003 | 0.811±0.028 | 0.978±0.007 | 0.849±0.026 |
|  |  | RPSSM | 0.850±0.025 | 0.773±0.033 | 0.986±0.002 | 0.809±0.025 | 0.974±0.007 | 0.835±0.022 |
|  |  | PSSM-AC | 0.854±0.026 | 0.786±0.035 | 0.986±0.003 | 0.818±0.024 | 0.978±0.006 | 0.853±0.034 |
|  | C2h | AAC-PSSM | 0.833±0.015 | 0.764±0.041 | 0.985±0.001 | 0.797±0.026 | 0.942±0.007 | 0.762±0.028 |
|  |  | DPC-PSSM | 0.830±0.013 | 0.775±0.034 | 0.984±0.002 | 0.801±0.022 | 0.961±0.007 | 0.817±0.019 |
|  |  | RPSSM | 0.829±0.017 | 0.775±0.032 | 0.984±0.002 | 0.801±0.021 | 0.958±0.007 | 0.806±0.019 |
|  |  | PSSM-AC | 0.835±0.014 | 0.780±0.030 | 0.985±0.002 | 0.807±0.021 | 0.960±0.007 | 0.813±0.019 |
|  | C2v | AAC-PSSM | 0.507±0.011 | 0.375±0.023 | 0.964±0.002 | 0.431±0.015 | 0.776±0.037 | 0.359±0.026 |
|  |  | DPC-PSSM | 0.617±0.325 | 0.062±0.065 | 0.989±0.016 | 0.098±0.074 | 0.719±0.063 | 0.340±0.108 |
|  |  | RPSSM | 0.586±0.085 | 0.346±0.037 | 0.974±0.015 | 0.431±0.042 | 0.745±0.045 | 0.378±0.055 |
|  |  | PSSM-AC | 0.554±0.205 | 0.218±0.138 | 0.976±0.028 | 0.278±0.146 | 0.808±0.056 | 0.388±0.111 |
|  | C3 | AAC-PSSM | 0.503±0.030 | 0.369±0.038 | 0.964±0.004 | 0.425±0.033 | 0.747±0.041 | 0.313±0.037 |
|  |  | DPC-PSSM | 0.045±0.091 | 0.017±0.042 | 0.990±0.019 | 0.023±0.057 | 0.674±0.059 | 0.292±0.101 |
|  |  | RPSSM | 0.554±0.118 | 0.311±0.051 | 0.970±0.024 | 0.389±0.059 | 0.715±0.045 | 0.334±0.054 |
|  |  | PSSM-AC | 0.401±0.330 | 0.144±0.133 | 0.974±0.038 | 0.174±0.148 | 0.744±0.044 | 0.311±0.082 |
| Protein embedding | C1 | EsmMean | 0.856±0.026 | 0.786±0.039 | 0.987±0.002 | 0.819±0.028 | 0.982±0.006 | 0.858±0.028 |
|  |  | ProtTrans | 0.856±0.024 | 0.793±0.039 | 0.987±0.002 | 0.823±0.028 | 0.981±0.007 | 0.849±0.031 |
|  |  | Doc2vec | 0.852±0.024 | 0.763±0.043 | 0.987±0.002 | 0.805±0.027 | 0.970±0.009 | 0.822±0.039 |
|  | C2h | EsmMean | 0.838±0.012 | 0.777±0.036 | 0.985±0.001 | 0.806±0.023 | 0.964±0.006 | 0.825±0.023 |
|  |  | ProtTrans | 0.837±0.013 | 0.783±0.035 | 0.985±0.002 | 0.808±0.022 | 0.964±0.005 | 0.827±0.019 |
|  |  | Doc2vec | 0.833±0.015 | 0.770±0.037 | 0.985±0.002 | 0.800±0.024 | 0.953±0.007 | 0.802±0.023 |
|  | C2v | EsmMean | 0.648±0.144 | 0.320±0.067 | 0.976±0.024 | 0.414±0.056 | 0.824±0.028 | 0.474±0.061 |
|  |  | ProtTrans | 0.703±0.272 | 0.102±0.131 | 0.983±0.033 | 0.125±0.114 | 0.822±0.033 | 0.390±0.058 |
|  |  | Doc2vec | 0.133±0.306 | 0.000±0.001 | 1.000±0.000 | 0.001±0.002 | 0.660±0.055 | 0.227±0.062 |
|  | C3 | EsmMean | 0.640±0.200 | 0.227±0.116 | 0.974±0.036 | 0.296±0.117 | 0.773±0.034 | 0.410±0.061 |
|  |  | ProtTrans | 0.047±0.093 | 0.061±0.123 | 0.977±0.044 | 0.053±0.106 | 0.751±0.037 | 0.315±0.064 |
|  |  | Doc2vec | 0.000±0.000 | 0.000±0.000 | 1.000±0.000 | 0.000±0.000 | 0.664±0.069 | 0.250±0.079 |

**Supplementary Table 7. Performance of the combination of the optimal encoding schemes in C1, C2h, C2v and C3 test sets**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Test set** | **Feature combination** | **Precision** | **Recall** | **Specificity** | **F1** | **AUROC** | **AUPRC** |
| C1 | CKSAAP+CTD+RPSSM+EsmMean | 0.855±0.026 | 0.789±0.037 | 0.987±0.003 | 0.820±0.027 | 0.981±0.007 | 0.863±0.027 |
|  | CKSAAP+CTD+RPSSM | 0.855±0.024 | 0.781±0.035 | 0.987±0.002 | 0.816±0.025 | 0.980±0.006 | 0.856±0.028 |
|  | CKSAAP+CTD+EsmMean | 0.855±0.026 | 0.791±0.039 | 0.987±0.003 | 0.821±0.029 | 0.981±0.007 | 0.862±0.028 |
|  | CKSAAP+RPSSM+EsmMean | 0.855±0.026 | 0.788±0.038 | 0.987±0.003 | 0.819±0.028 | 0.982±0.007 | 0.862±0.031 |
|  | CTD+RPSSM+EsmMean | 0.854±0.028 | 0.786±0.038 | 0.987±0.003 | 0.818±0.029 | 0.981±0.007 | 0.860±0.030 |
|  | CKSAAP+CTD | 0.855±0.026 | 0.776±0.041 | 0.987±0.003 | 0.813±0.028 | 0.979±0.007 | 0.860±0.027 |
|  | CKSAAP+RPSSM | 0.855±0.026 | 0.777±0.035 | 0.987±0.003 | 0.814±0.025 | 0.980±0.006 | 0.855±0.032 |
|  | CKSAAP+EsmMean | 0.855±0.026 | 0.789±0.039 | 0.987±0.003 | 0.820±0.028 | 0.981±0.007 | 0.860±0.029 |
|  | CTD+RPSSM | 0.852±0.028 | 0.777±0.037 | 0.986±0.003 | 0.812±0.027 | 0.978±0.007 | 0.846±0.025 |
|  | CTD+EsmMean | 0.854±0.025 | 0.787±0.039 | 0.987±0.002 | 0.819±0.029 | 0.982±0.006 | 0.866±0.027 |
|  | RPSSM+EsmMean | 0.856±0.026 | 0.789±0.036 | 0.987±0.002 | 0.820±0.027 | 0.982±0.007 | 0.859±0.029 |
| C2h | CKSAAP+CTD+RPSSM+EsmMean | 0.837±0.013 | 0.777±0.036 | 0.985±0.001 | 0.805±0.023 | 0.963±0.005 | 0.824±0.021 |
|  | CKSAAP+CTD+RPSSM | 0.835±0.015 | 0.773±0.038 | 0.985±0.002 | 0.802±0.024 | 0.959±0.006 | 0.812±0.020 |
|  | CKSAAP+CTD+EsmMean | 0.837±0.012 | 0.777±0.036 | 0.985±0.001 | 0.805±0.023 | 0.963±0.005 | 0.824±0.019 |
|  | CKSAAP+RPSSM+EsmMean | 0.837±0.013 | 0.777±0.036 | 0.985±0.001 | 0.805±0.023 | 0.963±0.006 | 0.826±0.022 |
|  | CTD+RPSSM+EsmMean | 0.838±0.013 | 0.780±0.034 | 0.985±0.001 | 0.808±0.021 | 0.963±0.005 | 0.823±0.021 |
|  | CKSAAP+CTD | 0.835±0.016 | 0.769±0.040 | 0.985±0.002 | 0.800±0.026 | 0.957±0.006 | 0.811±0.022 |
|  | CKSAAP+RPSSM | 0.836±0.015 | 0.771±0.038 | 0.985±0.002 | 0.802±0.024 | 0.958±0.006 | 0.813±0.020 |
|  | CKSAAP+EsmMean | 0.838±0.013 | 0.776±0.036 | 0.985±0.001 | 0.806±0.023 | 0.964±0.005 | 0.823±0.020 |
|  | CTD+RPSSM | 0.832±0.017 | 0.776±0.036 | 0.984±0.002 | 0.803±0.023 | 0.958±0.006 | 0.810±0.022 |
|  | CTD+EsmMean | 0.838±0.013 | 0.778±0.036 | 0.985±0.001 | 0.806±0.023 | 0.963±0.005 | 0.824±0.022 |
|  | RPSSM+EsmMean | 0.837±0.013 | 0.779±0.035 | 0.985±0.001 | 0.807±0.023 | 0.964±0.005 | 0.824±0.022 |
| C2v | CKSAAP+CTD+RPSSM+EsmMean | 0.643±0.147 | 0.321±0.055 | 0.975±0.026 | 0.414±0.041 | 0.827±0.031 | 0.477±0.063 |
|  | CKSAAP+CTD+RPSSM | 0.547±0.098 | 0.380±0.030 | 0.964±0.022 | 0.444±0.042 | 0.782±0.041 | 0.432±0.046 |
|  | CKSAAP+CTD+EsmMean | 0.659±0.156 | 0.302±0.063 | 0.976±0.027 | 0.397±0.045 | 0.820±0.027 | 0.470±0.061 |
|  | CKSAAP+RPSSM+EsmMean | 0.661±0.158 | 0.298±0.070 | 0.977±0.026 | 0.392±0.053 | 0.818±0.031 | 0.469±0.061 |
|  | CTD+RPSSM+EsmMean | 0.675±0.171 | 0.276±0.096 | 0.978±0.027 | 0.363±0.084 | 0.828±0.028 | 0.477±0.060 |
|  | CKSAAP+CTD | 0.557±0.103 | 0.359±0.041 | 0.967±0.023 | 0.429±0.040 | 0.784±0.035 | 0.431±0.042 |
|  | CKSAAP+RPSSM | 0.566±0.092 | 0.359±0.032 | 0.969±0.019 | 0.434±0.034 | 0.772±0.037 | 0.427±0.044 |
|  | CKSAAP+EsmMean | 0.660±0.149 | 0.295±0.077 | 0.978±0.024 | 0.389±0.064 | 0.825±0.032 | 0.475±0.061 |
|  | CTD+RPSSM | 0.543±0.095 | 0.378±0.028 | 0.964±0.022 | 0.441±0.039 | 0.814±0.050 | 0.452±0.047 |
|  | CTD+EsmMean | 0.707±0.193 | 0.244±0.108 | 0.980±0.027 | 0.328±0.096 | 0.829±0.029 | 0.475±0.060 |
|  | RPSSM+EsmMean | 0.706±0.184 | 0.244±0.099 | 0.981±0.026 | 0.332±0.084 | 0.824±0.030 | 0.472±0.059 |
| C3 | CKSAAP+CTD+RPSSM+EsmMean | 0.675±0.236 | 0.235±0.107 | 0.973±0.038 | 0.303±0.108 | 0.776±0.032 | 0.418±0.061 |
|  | CKSAAP+CTD+RPSSM | 0.489±0.111 | 0.367±0.032 | 0.955±0.030 | 0.413±0.057 | 0.727±0.031 | 0.375±0.050 |
|  | CKSAAP+CTD+EsmMean | 0.689±0.224 | 0.209±0.106 | 0.975±0.040 | 0.277±0.097 | 0.763±0.033 | 0.408±0.061 |
|  | CKSAAP+RPSSM+EsmMean | 0.697±0.234 | 0.203±0.112 | 0.975±0.038 | 0.267±0.105 | 0.769±0.034 | 0.409±0.058 |
|  | CTD+RPSSM+EsmMean | 0.709±0.254 | 0.194±0.117 | 0.975±0.039 | 0.254±0.113 | 0.775±0.034 | 0.411±0.061 |
|  | CKSAAP+CTD | 0.513±0.128 | 0.336±0.058 | 0.960±0.033 | 0.393±0.060 | 0.729±0.037 | 0.370±0.041 |
|  | CKSAAP+RPSSM | 0.538±0.110 | 0.343±0.051 | 0.965±0.027 | 0.411±0.057 | 0.731±0.031 | 0.375±0.043 |
|  | CKSAAP+EsmMean | 0.681±0.219 | 0.203±0.107 | 0.977±0.033 | 0.271±0.105 | 0.776±0.034 | 0.413±0.056 |
|  | CTD+RPSSM | 0.487±0.108 | 0.367±0.032 | 0.955±0.029 | 0.412±0.056 | 0.760±0.045 | 0.379±0.052 |
|  | CTD+EsmMean | 0.722±0.261 | 0.161±0.123 | 0.978±0.037 | 0.210±0.118 | 0.771±0.040 | 0.411±0.056 |
|  | RPSSM+EsmMean | 0.708±0.263 | 0.143±0.115 | 0.979±0.037 | 0.189±0.098 | 0.774±0.037 | 0.409±0.059 |

**Supplementary Table 8. Performance comparison of different machine learning methods based on final fusion feature encoding**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Test set** | **ML** | **Precision** | **Recall** | **Specificity** | **F1** | **AUROC** | **AUPRC** |
| C1 | RF | 0.855±0.026 | 0.789±0.037 | 0.987±0.003 | 0.820±0.027 | 0.981±0.007 | 0.863±0.027 |
|  | XGB | 0.853±0.025 | 0.791±0.039 | 0.986±0.002 | 0.820±0.029 | 0.977±0.008 | 0.850±0.030 |
|  | MLP | 0.790±0.061 | 0.754±0.045 | 0.979±0.009 | 0.769±0.033 | 0.951±0.016 | 0.770±0.055 |
|  | AB | 0.852±0.027 | 0.755±0.048 | 0.987±0.003 | 0.800±0.030 | 0.950±0.013 | 0.747±0.033 |
|  | SVM | 0.579±0.122 | 0.066±0.036 | 0.996±0.002 | 0.116±0.058 | 0.801±0.026 | 0.447±0.036 |
| C2h | RF | 0.837±0.013 | 0.777±0.036 | 0.985±0.001 | 0.805±0.023 | 0.963±0.005 | 0.824±0.021 |
|  | XGB | 0.839±0.013 | 0.783±0.033 | 0.985±0.002 | 0.810±0.021 | 0.959±0.007 | 0.817±0.020 |
|  | MLP | 0.782±0.066 | 0.760±0.043 | 0.978±0.008 | 0.770±0.044 | 0.944±0.010 | 0.760±0.040 |
|  | AB | 0.835±0.014 | 0.763±0.042 | 0.985±0.001 | 0.797±0.026 | 0.946±0.006 | 0.733±0.020 |
|  | SVM | 0.815±0.056 | 0.047±0.037 | 0.999±0.001 | 0.086±0.062 | 0.879±0.016 | 0.535±0.043 |
| C2v | RF | 0.643±0.147 | 0.321±0.055 | 0.975±0.026 | 0.414±0.041 | 0.827±0.031 | 0.477±0.063 |
|  | XGB | 0.771±0.140 | 0.151±0.117 | 0.992±0.009 | 0.223±0.143 | 0.781±0.048 | 0.370±0.080 |
|  | MLP | 0.424±0.112 | 0.421±0.043 | 0.936±0.028 | 0.416±0.064 | 0.693±0.054 | 0.331±0.090 |
|  | AB | 0.585±0.386 | 0.188±0.177 | 0.978±0.033 | 0.223±0.180 | 0.741±0.055 | 0.343±0.047 |
|  | SVM | 0.651±0.096 | 0.228±0.020 | 0.986±0.008 | 0.335±0.023 | 0.680±0.029 | 0.331±0.045 |
| C3 | RF | 0.675±0.236 | 0.235±0.107 | 0.973±0.038 | 0.303±0.108 | 0.776±0.032 | 0.418±0.061 |
|  | XGB | 0.333±0.356 | 0.093±0.120 | 0.994±0.008 | 0.134±0.160 | 0.731±0.055 | 0.333±0.077 |
|  | MLP | 0.413±0.120 | 0.414±0.037 | 0.933±0.030 | 0.406±0.065 | 0.686±0.041 | 0.322±0.087 |
|  | AB | 0.587±0.387 | 0.179±0.176 | 0.978±0.035 | 0.212±0.184 | 0.731±0.060 | 0.333±0.046 |
|  | SVM | 0.672±0.318 | 0.015±0.012 | 0.997±0.005 | 0.029±0.021 | 0.645±0.033 | 0.226±0.038 |

**Supplementary Table 9. Performance of different PPI prediction methods in the C1, C2h, C2v and C3 test sets and five-fold cross-validation**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Test set** | **Method** | **Precision** | **Recall** | **Specificity** | **F1** | **AUROC** | **AUPRC** |
| 5Fold | ML | 0.823±0.017 | 0.656±0.020 | 0.986±0.001 | 0.730±0.018 | 0.970±0.002 | 0.807±0.009 |
|  | IM | 0.428±0.026 | 0.441±0.012 | 0.941±0.005 | 0.434±0.019 | 0.694±0.009 | 0.284±0.022 |
|  | DDI | 0.333±0.013 | 0.587±0.017 | 0.882±0.007 | 0.425±0.012 | 0.763±0.007 | 0.412±0.029 |
|  | DMI | 0.119±0.003 | 0.677±0.017 | 0.498±0.006 | 0.202±0.004 | 0.628±0.013 | 0.176±0.009 |
|  | S | 0.146±0.008 | 0.544±0.028 | 0.683±0.006 | 0.231±0.012 | 0.643±0.015 | 0.161±0.007 |
| C1 | ML | 0.855±0.026 | 0.789±0.037 | 0.987±0.003 | 0.820±0.027 | 0.981±0.007 | 0.863±0.027 |
|  | IM | 0.567±0.106 | 0.377±0.041 | 0.969±0.011 | 0.450±0.050 | 0.675±0.021 | 0.291±0.042 |
|  | DDI | 0.514±0.101 | 0.584±0.046 | 0.940±0.023 | 0.542±0.062 | 0.779±0.021 | 0.496±0.056 |
|  | DMI | 0.117±0.012 | 0.638±0.040 | 0.512±0.028 | 0.197±0.018 | 0.622±0.034 | 0.205±0.025 |
|  | S | 0.137±0.011 | 0.460±0.053 | 0.708±0.025 | 0.210±0.017 | 0.608±0.024 | 0.151±0.018 |
| C2h | ML | 0.837±0.013 | 0.777±0.036 | 0.985±0.001 | 0.805±0.023 | 0.963±0.005 | 0.824±0.021 |
|  | IM | 0.543±0.062 | 0.305±0.021 | 0.974±0.007 | 0.389±0.021 | 0.640±0.010 | 0.254±0.026 |
|  | DDI | 0.503±0.070 | 0.524±0.034 | 0.946±0.019 | 0.509±0.034 | 0.745±0.011 | 0.435±0.038 |
|  | DMI | 0.113±0.005 | 0.582±0.026 | 0.546±0.024 | 0.189±0.008 | 0.604±0.011 | 0.186±0.007 |
|  | S | 0.129±0.007 | 0.462±0.025 | 0.690±0.024 | 0.202±0.009 | 0.602±0.012 | 0.145±0.009 |
| C2v | ML | 0.643±0.147 | 0.321±0.055 | 0.975±0.026 | 0.414±0.041 | 0.827±0.031 | 0.477±0.063 |
|  | IM | 0.438±0.058 | 0.351±0.023 | 0.954±0.010 | 0.388±0.029 | 0.653±0.013 | 0.223±0.026 |
|  | DDI | 0.363±0.054 | 0.411±0.032 | 0.926±0.016 | 0.384±0.040 | 0.674±0.018 | 0.245±0.042 |
|  | DMI | 0.111±0.012 | 0.654±0.016 | 0.475±0.034 | 0.189±0.017 | 0.581±0.018 | 0.128±0.015 |
|  | S | 0.145±0.019 | 0.523±0.077 | 0.692±0.049 | 0.226±0.023 | 0.624±0.029 | 0.138±0.014 |
| C3 | ML | 0.675±0.236 | 0.235±0.107 | 0.973±0.038 | 0.303±0.108 | 0.776±0.032 | 0.418±0.061 |
|  | IM | 0.345±0.068 | 0.180±0.025 | 0.964±0.010 | 0.235±0.033 | 0.572±0.014 | 0.152±0.019 |
|  | DDI | 0.319±0.068 | 0.300±0.035 | 0.933±0.019 | 0.307±0.045 | 0.618±0.021 | 0.193±0.039 |
|  | DMI | 0.103±0.012 | 0.550±0.042 | 0.518±0.035 | 0.173±0.017 | 0.550±0.023 | 0.125±0.014 |
|  | S | 0.126±0.021 | 0.436±0.067 | 0.692±0.056 | 0.193±0.025 | 0.574±0.028 | 0.119±0.014 |

**Supplementary Table 10. Distribution of predicted positive Samples by different methods**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Test set** |  | **Total** | **FN** | **TP** | **ML** | **IM** | **DDI** | **DMI** | **S** |
| 5fold | 1 | 405 | 33 | 378 | 279 | 172 | 232 | 267 | 222 |
|  | 2 | 404 | 41 | 368 | 273 | 181 | 246 | 282 | 207 |
|  | 3 | 404 | 37 | 372 | 262 | 186 | 241 | 287 | 234 |
|  | 4 | 404 | 36 | 373 | 258 | 185 | 250 | 284 | 215 |
|  | 5 | 404 | 34 | 376 | 275 | 180 | 235 | 275 | 237 |
|  | Avg | 404 | 36 | 373 | 269 | 181 | 241 | 279 | 223 |
| C1 | 1 | 99 | 8 | 91 | 76 | 39 | 59 | 66 | 44 |
|  | 2 | 93 | 7 | 86 | 79 | 40 | 62 | 64 | 40 |
|  | 3 | 99 | 10 | 89 | 77 | 35 | 53 | 60 | 39 |
|  | 4 | 100 | 11 | 89 | 87 | 40 | 52 | 61 | 52 |
|  | 5 | 99 | 4 | 95 | 86 | 38 | 59 | 63 | 50 |
|  | 6 | 100 | 5 | 95 | 82 | 42 | 67 | 70 | 48 |
|  | 7 | 96 | 2 | 94 | 77 | 30 | 57 | 60 | 50 |
|  | 8 | 100 | 12 | 88 | 77 | 48 | 64 | 71 | 59 |
|  | 9 | 100 | 6 | 94 | 80 | 36 | 62 | 73 | 41 |
|  | 10 | 98 | 6 | 92 | 79 | 35 | 57 | 62 | 44 |
|  | Avg | 98 | 7 | 91 | 80 | 38 | 59 | 65 | 47 |
| C2h | 1 | 428 | 21 | 407 | 333 | 131 | 233 | 254 | 206 |
|  | 2 | 396 | 28 | 368 | 335 | 119 | 205 | 228 | 175 |
|  | 3 | 437 | 43 | 394 | 338 | 129 | 222 | 253 | 205 |
|  | 4 | 445 | 45 | 400 | 316 | 132 | 210 | 260 | 181 |
|  | 5 | 430 | 26 | 404 | 337 | 124 | 214 | 238 | 210 |
|  | 6 | 429 | 27 | 402 | 332 | 152 | 260 | 279 | 214 |
|  | 7 | 428 | 32 | 396 | 324 | 127 | 227 | 249 | 197 |
|  | 8 | 432 | 39 | 393 | 330 | 127 | 222 | 251 | 201 |
|  | 9 | 426 | 30 | 396 | 341 | 141 | 225 | 249 | 189 |
|  | 10 | 429 | 37 | 392 | 324 | 119 | 215 | 235 | 193 |
|  | Avg | 428 | 33 | 395 | 331 | 130 | 223 | 250 | 197 |
| C2v | 1 | 405 | 102 | 303 | 94 | 146 | 181 | 271 | 183 |
|  | 2 | 497 | 94 | 403 | 170 | 160 | 173 | 318 | 249 |
|  | 3 | 422 | 80 | 342 | 119 | 159 | 188 | 281 | 258 |
|  | 4 | 399 | 53 | 346 | 146 | 159 | 180 | 266 | 271 |
|  | 5 | 433 | 104 | 329 | 139 | 148 | 173 | 290 | 194 |
|  | 6 | 417 | 82 | 335 | 150 | 134 | 159 | 272 | 200 |
|  | 7 | 459 | 87 | 372 | 199 | 158 | 187 | 315 | 200 |
|  | 8 | 396 | 82 | 314 | 133 | 136 | 153 | 252 | 200 |
|  | 9 | 399 | 71 | 328 | 114 | 136 | 165 | 253 | 240 |
|  | 10 | 433 | 87 | 346 | 117 | 159 | 191 | 294 | 227 |
|  | Avg | 426 | 84 | 342 | 138 | 150 | 175 | 281 | 222 |
| C3 | 1 | 198 | 53 | 145 | 15 | 40 | 71 | 125 | 74 |
|  | 2 | 212 | 38 | 174 | 68 | 27 | 45 | 113 | 97 |
|  | 3 | 181 | 49 | 132 | 14 | 32 | 54 | 85 | 98 |
|  | 4 | 162 | 29 | 133 | 44 | 27 | 52 | 94 | 87 |
|  | 5 | 178 | 39 | 139 | 55 | 37 | 51 | 98 | 66 |
|  | 6 | 173 | 36 | 137 | 52 | 26 | 51 | 99 | 61 |
|  | 7 | 182 | 43 | 139 | 70 | 33 | 56 | 104 | 70 |
|  | 8 | 176 | 45 | 131 | 44 | 33 | 51 | 92 | 74 |
|  | 9 | 175 | 29 | 146 | 47 | 36 | 55 | 101 | 88 |
|  | 10 | 190 | 55 | 135 | 19 | 37 | 61 | 100 | 79 |
|  | Avg | 183 | 42 | 141 | 43 | 33 | 55 | 101 | 79 |

**Supplementary Table 11. Performance of integrating five different methods based on different meta learners**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Test set** | **Meta learner** | **Precision** | **Recall** | **Specificity** | **F1** | **AUROC** | **AUPRC** |
| 5fold | RF | 0.813±0.018 | 0.722±0.023 | 0.983±0.002 | 0.764±0.013 | 0.972±0.003 | 0.826±0.015 |
|  | LR | 0.690±0.016 | 0.827±0.012 | 0.963±0.003 | 0.752±0.012 | 0.969±0.003 | 0.811±0.012 |
|  | SVM | 0.812±0.014 | 0.684±0.019 | 0.984±0.001 | 0.742±0.016 | 0.968±0.003 | 0.810±0.012 |
|  | NB | 0.534±0.021 | 0.803±0.012 | 0.930±0.006 | 0.641±0.016 | 0.935±0.004 | 0.761±0.004 |
|  | CKa | 0.823±0.017 | 0.656±0.020 | 0.986±0.001 | 0.730±0.018 | 0.970±0.002 | 0.807±0.009 |
| C1 | RF | 0.873±0.020 | 0.812±0.031 | 0.988±0.002 | 0.841±0.022 | 0.983±0.008 | 0.894±0.020 |
|  | LR | 0.811±0.025 | 0.844±0.028 | 0.980±0.003 | 0.827±0.024 | 0.979±0.007 | 0.861±0.028 |
|  | SVM | 0.850±0.026 | 0.801±0.033 | 0.986±0.003 | 0.825±0.027 | 0.978±0.007 | 0.859±0.024 |
|  | NB | 0.628±0.091 | 0.853±0.024 | 0.947±0.019 | 0.720±0.064 | 0.952±0.011 | 0.817±0.026 |
|  | CKa | 0.855±0.026 | 0.789±0.037 | 0.987±0.003 | 0.820±0.027 | 0.981±0.007 | 0.863±0.027 |
| C2h | RF | 0.857±0.018 | 0.798±0.029 | 0.987±0.002 | 0.826±0.020 | 0.975±0.004 | 0.872±0.017 |
|  | LR | 0.782±0.023 | 0.834±0.027 | 0.977±0.004 | 0.807±0.019 | 0.960±0.007 | 0.819±0.020 |
|  | SVM | 0.835±0.014 | 0.784±0.034 | 0.985±0.002 | 0.809±0.022 | 0.962±0.007 | 0.815±0.020 |
|  | NB | 0.624±0.065 | 0.835±0.029 | 0.948±0.017 | 0.712±0.046 | 0.937±0.009 | 0.769±0.024 |
|  | CKa | 0.837±0.013 | 0.777±0.036 | 0.985±0.001 | 0.805±0.023 | 0.963±0.005 | 0.824±0.021 |
| C2v | RF | 0.766±0.089 | 0.400±0.035 | 0.987±0.008 | 0.525±0.046 | 0.886±0.030 | 0.614±0.065 |
|  | LR | 0.421±0.085 | 0.542±0.051 | 0.919±0.035 | 0.469±0.062 | 0.817±0.029 | 0.477±0.067 |
|  | SVM | 0.666±0.117 | 0.324±0.041 | 0.980±0.017 | 0.429±0.036 | 0.826±0.026 | 0.474±0.063 |
|  | NB | 0.407±0.066 | 0.494±0.031 | 0.925±0.022 | 0.444±0.043 | 0.762±0.035 | 0.429±0.055 |
|  | CKa | 0.643±0.147 | 0.321±0.055 | 0.975±0.026 | 0.414±0.041 | 0.827±0.031 | 0.477±0.063 |
| C3 | RF | 0.687±0.144 | 0.253±0.039 | 0.985±0.014 | 0.362±0.040 | 0.846±0.033 | 0.506±0.075 |
|  | LR | 0.375±0.104 | 0.506±0.062 | 0.900±0.056 | 0.418±0.064 | 0.771±0.026 | 0.394±0.058 |
|  | SVM | 0.551±0.093 | 0.159±0.054 | 0.984±0.014 | 0.237±0.044 | 0.775±0.025 | 0.374±0.058 |
|  | NB | 0.391±0.090 | 0.447±0.042 | 0.924±0.029 | 0.411±0.054 | 0.729±0.027 | 0.325±0.042 |
|  | CKa | 0.675±0.236 | 0.235±0.107 | 0.973±0.038 | 0.303±0.108 | 0.776±0.032 | 0.418±0.061 |
| a. CK represents the ML method with relatively better performance among the five prediction methods | | | | | | |