**Table 1**. Functional groups within 508 compound diverse dataset

|  |  |
| --- | --- |
| **Chemical class** | **No. of compounds** |
| Aliphatic alkanes, alkenes, alkynes | 124 |
| Monocyclic compounds | 260 |
| Monocyclic carbocycles | 186 |
| Monocyclic heterocycles | 74 |
| Polycyclic compounds | 192 |
| Polycyclic carbocycles | 119 |
| Polycyclic heterocycles | 73 |
| Nitro compounds | 47 |
| Nitroso compounds | 30 |
| Alkyl halides | 55 |
| Alcohols, thiols | 93 |
| Ethers, sulfides | 38 |
| Ketones, ketenes, imines, quinones | 39 |
| Carboxylic acids, peroxy acids | 34 |
| Esters, lactones | 34 |
| Amides, imides, lactams | 36 |
| Carbamates, ureas, thioureas, guanidines | 41 |
| Amines, hydroxylamines | 143 |
| Hydrazines, hydrazides, hydrazones, traizines | 55 |
| Oxygenated sulfur and phosphorus | 53 |
| Epoxides, peroxides, aziridines | 25 |

**Table 2.** Number of different types of variables in the two datasets

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Dataset** | **Number of variables for each type** | | | |
| **TS** | **TC** | **3-D** | **QC** |
| **508 compound** | 103 | 195 | 3 | 6 |
| **95 amines** | 108 | 158 | 3 | 6 |

**Table 3.** Comparison of performances of model 1 (RR and ITC+RR) for diverse and congeneric datasets

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Dataset used** | **Predictive model** | **Type of predictor used** | **No. of predictors** | **Correct classification %** | **Sensitivity** | **Specificity** |
| **508 compound diverse dataset** | Ridge regression  [33] | TS | 103 | 53.14 | 52.34 | 53.97 |
| TS+TC | 298 | 76.97 | 83.98 | 69.84 |
| TS+TC+3D+QC | 307 | 77.17 | 84.38 | 69.84 |
| ITC+ RR | TS | 103 | 66.34 | 73.83 | 58.73 |
| TS+TC | 298 | 73.23 | 77.34 | 69.05 |
| TS+TC+3D | 301 | 74.80 | 77.34 | 72.22 |
| TS+TC+3D+QC | 307 | 72.05 | 76.17 | 67.86 |
| TS+TC+AP [27] | 2620 | 78.35 | 84.38 | 72.22 |
| **95 amines congeneric dataset** | Ridge regression | TS | 108 | 83.16 | 75.47 | 88.42 |
| TS+TC | 266 | 84.21 | 77.36 | 92.86 |
| TS+TC+3D | 269 | 84.21 | 77.36 | 92.86 |
| TS+TC+3D+QC | 275 | 84.21 | 77.36 | 92.86 |
| ITC + RR | TS | 108 | 88.42 | 92.45 | 83.33 |
| TS+TC | 266 | 89.47 | 92.45 | 85.71 |
| TS+TC+3D | 269 | 88.42 | 92.45 | 83.33 |
| TS+TC+3D+QC | 275 | 85.26 | 88.68 | 80.95 |

**Table 4.** Comparison of performances of model 2 (Envelope LDA) diverse and congeneric datasets

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Dataset used** | **Type of predictor used** | **No. predictors** | **Correct classification %** | **Sensitivity** | **Specificity** |
| **508 compound diverse dataset** | TS | 103 | 57.09 | 65.63 | 48.41 |
| TS+TC | 298 | 60.24 | 69.92 | 46.43 |
| TS+TC+3D | 301 | 61.02 | 71.09 | 50.79 |
| TS+TC+3D+QC | 307 | 64.37 | 69.14 | 59.52 |
| **95 amines congeneric dataset** | TS | 108 | 81.05 | 92.86 | 71.70 |
| TS+TC | 266 | 80.00 | 83.33 | 77.36 |
| TS+TC+3D | 269 | 80.00 | 83.33 | 77.36 |
| TS+TC+3D+QC | 275 | 71.58 | 78.57 | 66.04 |