Drug Components-Disease Network Related to Acute Lung Injury Inference Based on Forest Graph-embedded Deep Feedforward Network

Bin Yang  
Zaozhuang University

Wenzheng Bao (✉ baowz55555@126.com )  
Xuzhou University of Technology

Jinglong Wang  
Zaozhuang University

Baitong Chen  
Xuzhou University of Technology

Research Article

Keywords: network pharmacology, acute lung injury, deep learning algorithm, target

Posted Date: January 27th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1238402/v1

License: © This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

**Background:** Acute lung injury (ALI) is a serious respiratory disease, which can lead to acute respiratory failure or death. It is closely related to the pathogenesis of New Coronavirus pneumonia (COVID-19). Many researches showed that traditional Chinese medicine (TCM) had a good effect on its intervention, and network pharmacology could play a very important role.

**Results:** In order to construct "disease-gene-target-drug" interaction network more accurately, deep learning algorithm is utilized in this paper. Two ALI-related target genes (REAL and SATA3) are considered, and the active and inactive compounds of the two corresponding target genes are collected as training data, respectively. Molecular descriptors and molecular fingerprints are utilized to characterize each compound. Forest graph embedded deep feed forward network (forgeNet) is proposed to train and identify 19 compounds in Erhuang decoction (EhD) and Dexamethasone (DXMS).

**Conclusions:** The experiment results show that forgeNet performs better than support vector machines (SVM), random forest (RF) and gcForest.

1. **Background**

Internal and external etiology can lead to self-stable regulation disorder, which could change a series of metabolisms, functions and structures. Abnormal life activity processes are manifested as abnormal symptoms, signs and behavior [1–2]. Under certain conditions, the abnormal life activity processes caused by the disturbance of homeostasis after the damage of the disease cause the disease [3–4]. Traditional Chinese medicine (TCM) has been utilized to treat diseases for thousands of years [5–7]. Traditional Chinese medicine is a kind of material with the function of rehabilitation and health care, which could be utilized to prevent, treat and diagnose diseases under the guidance of TCM theory [8–11].

Traditional Chinese medicine mainly comes from natural medicine and its processed products, including plant medicine, animal medicine, mineral medicine and some chemical and biological products [12–13]. The most important feature of traditional Chinese medicine in treating diseases is to pay attention to the adjustment of the functions of viscera and organs, and the balance and coordination between them. The focus of traditional Chinese medicine treatment is not that the human body is infected with the specific bacteria, virus and other pathogenic factors, but the specific reaction of the human body after these pathogenic factors act on the human body [14–15]. The purpose of treatment is to enhance the disease resistance and recovery ability of human body. To kill bacteria and relieve symptoms are mainly achieved by enhancing the body's own functions. In recent years, traditional Chinese medicine has certain advantages in the treatment of pneumonia [16], shock [17], convulsion [18], hemorrhage [19], acute respiratory failure [20], renal failure [21], heart failure [22], cerebrovascular accident [23], etc. it is not only effective, but also safe and simple, with few adverse reactions.

In the past decade, with the rapid development of sequencing technology, a large number of genomics data such as genomics, proteomics, metabonomics and so on, have been generated, which has led to the
changes in the research of traditional Chinese medicine for diseases. Network pharmacology has been proposed, which was developed on the basis of the rapid development of systems biology and computer technology, generating the "disease-gene-target-drug" interaction network. Through network analysis, we can systematically and comprehensively observe the intervention and influence of drugs on the disease network, reveal the mystery of the synergistic effect of multi branch drugs on the human body, and find out the multi-target new drugs with high efficiency and low toxicity. Network pharmacology of traditional Chinese medicine has become a new idea for drug mechanism research and new drug development [24–28]. Lu et al. utilized network pharmacology and molecular docking technology to study the mechanism of Shaoyao Decoction in the treatment of ulcerative colitis, and found that Shaoyao decoction can improve the pathological damage of colon [29]. Liu et al. collected the main active components of Portulacae Herba, constructed interaction network of target proteins of liver cancer, and found that ketones may be the main material basis of its anti-liver cancer, which is related to the regulation of MAPK signaling pathway [30]. Liu et al. utilized network pharmacology to screen 102 active components of Danzhi Xiaoyao Powder, 147 corresponding targets and 52 intersecting targets with insomnia, and obtained the key components, key targets and key pathways of Danzhi Xiaoyao Powder in the treatment of insomnia [31]. Yang et al. presented network pharmacology to analyze the potential anti-tumor mechanisms of the main active components of Prunella vulgaris systematically at the molecular level [32]. Shen et al. discussed the possible mechanism of Wuling Powder in the treatment of diabetic nephropathy by network pharmacology, and found that Wuling Powder may reduce renal cell damage by regulating apoptosis related proteins, such as Caspases family protein and BCL2 Protein family [33].

In the recent years, data mining methods have been applied to extract useful information from lots of TCM data [33]. Ren et al. utilized data mining methods to screen out 47 prescriptions, and found out 14 core drugs and 7 new prescriptions in order to search the medication rules and mechanism of TCM in the treatment of carotid atherosclerosis (CAS) [34]. Ga et al. utilized data mining method to select the top five active components of each Tibetan medicine with high frequency and network pharmacology was utilized to analyze the mechanism of Tibetan medicine in the treatment of high altitude polycythemia [35]. In order to study the medication rule of TCM intervention in iron death, Ou et al. constructed target-compound, compound-TCM, target-compound-TCM network, and frequency statistics was utilized to show that bitter and pungent herbs were the main herbs that could interfere with iron death, while cold herbs were the main ones, which mainly belonged to liver and lung meridians [36]. Pan et al. reprocessed a large number of Chinese medicine prescriptions for the treatment of primary liver cancer, and by analysis of data mining and network pharmacology medication regularity of effective traditional Chinese medicine prescriptions in the treatment of primary liver cancer was obtained [37]. Zheng et al. presented four classifiers to infer compound-target interaction network in the process of network pharmacology analysis [38].

In order to better mine omics data and construct "disease-gene-target-drug" interaction network, deep learning model was utilized in this paper. Taking acute lung injury (ALI) disease as an example, we selected two disease-related target genes (REAL and SATA3). The active and inactive compounds of the two target genes combined are collected. Molecular descriptors and molecular fingerprints are utilized to
characterize each compound, which contain 374 features. Forest graph embedded deep feed forward network is utilized to train and identify new compounds target genes related.

2. Results

In this section, active and inactive ligands of two key target genes: REAL and SATA3 about ALI disease are collected. For REAL, 966 ligands are collected, which contain 146 positive samples and 820 negative samples (Data1). For SATA3, 193 active ligands and 1210 inactive ligands are collected (Data2). Molecular descriptors and molecular fingerprints of each ligand could be obtained, which contains 374 features. In order to better reflect the effectiveness of forgeNet, three classical classifiers (SVM [42], RF [43] and gcForest [44]) are utilized to identify the compounds associated with diseases. Five evaluation criteria of classifier performance are utilized, which are $SN$, $SP$, $ACC$, $MCC$ and $F1$, respectively.

2.1 Model test

In order to test the generalization and stability of forgeNet, leave-one-out, 3-fold, 5-fold and 10-fold cross validation methods are utilized. With Data1, the inference performances of four methods with leave-one-out, 3-fold, 5-fold and 10-fold cross validation methods are listed in Table 1, Table 3, Table 5 and Table 7, respectively. With Data2, the inference performances of four methods with leave-one-out, 3-fold, 5-fold and 10-fold cross validation methods are listed in Table 2, Table 4, Table 6 and Table 8, respectively. From the results, it could be seen that gcForest has the highest $SN$ performances among four methods, which reveal that gcForest could identify more true active ligands. RF could obtain the higher $SP$ than SVM, gcForest and forgeNet, which show that RF could identify true more inactive ligands. ForgeNet has the best $ACC$, $MCC$ and $F1$ performances among four methods. The results reveal that forgeNet could identify more true active and inactive ligands than SVM, RF and gcForest. And when two categories have very different sizes, forgeNet performs best. $F1$ performances show that on the whole forgeNet could infer components-disease network more accurately than other three classifiers.

| Table 1 | Performances of four methods with Data1 by leave-one-out method. |
|---|---|---|---|---|---|
| | $SN$ | $SP$ | $ACC$ | $MCC$ | $F1$ |
| SVM | 0.664384 | 0.968293 | 0.92236 | 0.679827 | 0.72119 |
| RF | 0.534247 | **0.99878** | 0.928571 | 0.696718 | 0.693333 |
| gcForest | **0.828767** | 0.960976 | 0.940994 | 0.774779 | 0.809365 |
| ForgeNet | 0.80137 | 0.981707 | **0.954451** | **0.816616** | **0.841727** |
Table 2
Performances of four methods with Data2 by leave-one-out method.

<table>
<thead>
<tr>
<th>Method</th>
<th>SN</th>
<th>SP</th>
<th>ACC</th>
<th>MCC</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>0.549223</td>
<td>0.975207</td>
<td>0.916607</td>
<td>0.610521</td>
<td>0.644377</td>
</tr>
<tr>
<td>RF</td>
<td>0.533679</td>
<td><strong>0.997521</strong></td>
<td>0.933713</td>
<td>0.692319</td>
<td>0.688963</td>
</tr>
<tr>
<td>gcForest</td>
<td>0.735751</td>
<td>0.952893</td>
<td>0.923022</td>
<td>0.679871</td>
<td>0.72449</td>
</tr>
<tr>
<td>ForgeNet</td>
<td><strong>0.751295</strong></td>
<td>0.978512</td>
<td><strong>0.947256</strong></td>
<td><strong>0.768381</strong></td>
<td><strong>0.796703</strong></td>
</tr>
</tbody>
</table>

Table 3
Performances of four methods with Data1 by 3-fold cross validation method.

<table>
<thead>
<tr>
<th>Method</th>
<th>SN</th>
<th>SP</th>
<th>ACC</th>
<th>MCC</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>0.589041</td>
<td>0.964634</td>
<td>0.907867</td>
<td>0.612385</td>
<td>0.659004</td>
</tr>
<tr>
<td>RF</td>
<td>0.356164</td>
<td><strong>0.995122</strong></td>
<td>0.898551</td>
<td>0.538431</td>
<td>0.514851</td>
</tr>
<tr>
<td>gcForest</td>
<td><strong>0.842466</strong></td>
<td>0.876829</td>
<td>0.871636</td>
<td>0.61047</td>
<td>0.664865</td>
</tr>
<tr>
<td>ForgeNet</td>
<td>0.69863</td>
<td>0.97561</td>
<td><strong>0.933747</strong></td>
<td><strong>0.727021</strong></td>
<td><strong>0.761194</strong></td>
</tr>
</tbody>
</table>

Table 4
Performances of four methods with Data2 by 3-fold cross validation method.

<table>
<thead>
<tr>
<th>Method</th>
<th>SN</th>
<th>SP</th>
<th>ACC</th>
<th>MCC</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>0.264249</td>
<td>0.980165</td>
<td>0.881682</td>
<td>0.374255</td>
<td>0.380597</td>
</tr>
<tr>
<td>RF</td>
<td>0.316062</td>
<td><strong>0.997521</strong></td>
<td>0.903778</td>
<td>0.517658</td>
<td>0.474708</td>
</tr>
<tr>
<td>gcForest</td>
<td><strong>0.580311</strong></td>
<td>0.704132</td>
<td>0.687099</td>
<td>0.207576</td>
<td>0.337858</td>
</tr>
<tr>
<td>ForgeNet</td>
<td>0.466321</td>
<td>0.975207</td>
<td><strong>0.905203</strong></td>
<td><strong>0.543783</strong></td>
<td><strong>0.57508</strong></td>
</tr>
</tbody>
</table>
Table 5
Performances of four methods with Data1 by 5-fold cross validation method.

<table>
<thead>
<tr>
<th>Method</th>
<th>SN</th>
<th>SP</th>
<th>ACC</th>
<th>MCC</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>0.60274</td>
<td>0.962195</td>
<td>0.907867</td>
<td>0.615696</td>
<td>0.664151</td>
</tr>
<tr>
<td>RF</td>
<td>0.39726</td>
<td><strong>0.997561</strong></td>
<td>0.906832</td>
<td>0.585928</td>
<td>0.563107</td>
</tr>
<tr>
<td>gcForest</td>
<td><strong>0.787671</strong></td>
<td>0.937805</td>
<td>0.915114</td>
<td>0.688823</td>
<td>0.737179</td>
</tr>
<tr>
<td>ForgeNet</td>
<td>0.691781</td>
<td>0.97439</td>
<td><strong>0.931677</strong></td>
<td><strong>0.71832</strong></td>
<td><strong>0.753731</strong></td>
</tr>
</tbody>
</table>

Table 6
Performances of four methods with Data2 by 5-fold cross validation method.

<table>
<thead>
<tr>
<th>Method</th>
<th>SN</th>
<th>SP</th>
<th>ACC</th>
<th>MCC</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>0.388601</td>
<td>0.957025</td>
<td>0.878831</td>
<td>0.414907</td>
<td>0.46875</td>
</tr>
<tr>
<td>RF</td>
<td>0.139896</td>
<td>1</td>
<td>0.881682</td>
<td>0.350741</td>
<td>0.245455</td>
</tr>
<tr>
<td>gcForest</td>
<td><strong>0.585492</strong></td>
<td>0.835537</td>
<td>0.80114</td>
<td>0.348734</td>
<td>0.447525</td>
</tr>
<tr>
<td>ForgeNet</td>
<td>0.502591</td>
<td>0.966942</td>
<td><strong>0.903065</strong></td>
<td><strong>0.544829</strong></td>
<td><strong>0.587879</strong></td>
</tr>
</tbody>
</table>

Table 7
Performances of four methods with Data1 by 10-fold cross validation method.

<table>
<thead>
<tr>
<th>Method</th>
<th>SN</th>
<th>SP</th>
<th>ACC</th>
<th>MCC</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>0.630137</td>
<td>0.968293</td>
<td>0.917184</td>
<td>0.654572</td>
<td>0.69697</td>
</tr>
<tr>
<td>RF</td>
<td>0.486301</td>
<td><strong>0.997561</strong></td>
<td>0.92029</td>
<td>0.655719</td>
<td>0.648402</td>
</tr>
<tr>
<td>gcForest</td>
<td><strong>0.80137</strong></td>
<td>0.947561</td>
<td>0.925466</td>
<td>0.721602</td>
<td>0.764706</td>
</tr>
<tr>
<td>ForgeNet</td>
<td>0.760274</td>
<td>0.980488</td>
<td><strong>0.947205</strong></td>
<td><strong>0.785198</strong></td>
<td><strong>0.813187</strong></td>
</tr>
</tbody>
</table>
Table 8
Performances of four methods with Data2 by 10-fold cross validation method.

<table>
<thead>
<tr>
<th></th>
<th>SN</th>
<th>SP</th>
<th>ACC</th>
<th>MCC</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>0.455959</td>
<td>0.961157</td>
<td>0.891661</td>
<td>0.487193</td>
<td>0.536585</td>
</tr>
<tr>
<td>RF</td>
<td>0.378238</td>
<td>0.999174</td>
<td>0.913756</td>
<td>0.581579</td>
<td>0.546816</td>
</tr>
<tr>
<td>gcForest</td>
<td>0.601036</td>
<td>0.928099</td>
<td>0.883108</td>
<td>0.518083</td>
<td>0.585859</td>
</tr>
<tr>
<td>ForgeNet</td>
<td>0.595855</td>
<td>0.973554</td>
<td>0.921597</td>
<td>0.640386</td>
<td>0.676471</td>
</tr>
</tbody>
</table>

Receiver operating characteristic (ROC) curve is based on false positive rate (FPR) and true positives rate (TPR). Area under curve (AUC) is defined as the area under the ROC curve surrounded by the coordinate axis. In general ROC and AUC are utilized to evaluate the classifiers. With Data1, ROC curve and AUC performances of four methods with 3-fold, 5-fold, 10-fold cross validation methods and leave-one-out are depicted in Fig. 1, Fig. 2, Fig. 3 and Fig. 4, respectively. With Data2, ROC curve and AUC performances of four methods with 3-fold, 5-fold, 10-fold cross validation methods and leave-one-out are depicted in Fig. 5, Fig. 6, Fig. 7 and Fig. 8, respectively. From the results, in most of cases, forgeNet could obtain the best AUC performances among four classifiers. With Data1 and Data2, RF performs best in terms of ROC and AUC by leave-one-out method.

2.2 Compound screening

19 active chemical compounds (Neoglycyrol, Uralenol, Syringic acid 4-β-D-Glucopyranoside, Gancaonin N, Chrysin -6-C-glucoside-8-C-arabinoside, Chrysin-6-C-arabinoside 8-C-glucoside, Liiquiritin, Baicalin, Isomer of Baicalin, Oroxylin A-7-O-β-D-glucuronide, Chrysin-7-O-glucuronide, Isoliquiritin, Wogonoside, Liquiritigenin, Baicalein, Isoliquiritigenin, Wogonin, Oroxylin A, and Glycyrrhetinic acid) in Erhuang decoction (EhD) can dock with ALI related target genes and have high potential biological activity, which have been proved in the reference [39]. Dexamethasone (DXMS) is used as control drug. Molecular descriptors and molecular fingerprints are also utilized to obtain the features of 20 chemical compounds. Data1 and Data2 are utilized as the training sets in order to predict 20 chemical compounds, respectively. The prediction ranks are listed in Table 9. By ranking results, we can see that DXMS ranks last by forgeNet on average, which is consistent with the results of molecular docking in the past research [39]. Thus the results reveal that forgeNet could screen the chemical compounds more accurately than SVM, RF and gcForest.
Table 9
Prediction ranks of 20 chemical compounds by SVM, RF, gcForest and forgeNet.

<table>
<thead>
<tr>
<th>ID</th>
<th>Chemical compounds</th>
<th>SVM</th>
<th>SVM</th>
<th>RF</th>
<th>RF</th>
<th>gcForest</th>
<th>gcForest</th>
<th>forgeNet</th>
<th>forgeNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Neoglycyrol</td>
<td>11</td>
<td>13</td>
<td>18</td>
<td>8</td>
<td>12</td>
<td>11</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Uralenol</td>
<td>12</td>
<td>15</td>
<td>12</td>
<td>6</td>
<td>16</td>
<td>6</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>Syringic acid 4-β-D-Glucopyranoside</td>
<td>3</td>
<td>4</td>
<td>7</td>
<td>9</td>
<td>17</td>
<td>11</td>
<td>1</td>
<td>17</td>
</tr>
<tr>
<td>4</td>
<td>Gancaonin N</td>
<td>13</td>
<td>16</td>
<td>19</td>
<td>10</td>
<td>15</td>
<td>10</td>
<td>12</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>Chrysin-6-C-glucoside-8-C-arabinoside</td>
<td>14</td>
<td>19</td>
<td>15</td>
<td>18</td>
<td>11</td>
<td>19</td>
<td>10</td>
<td>18</td>
</tr>
<tr>
<td>6</td>
<td>Chrysin-6-C-arabinoside 8-C-glucoside</td>
<td>15</td>
<td>18</td>
<td>14</td>
<td>19</td>
<td>9</td>
<td>20</td>
<td>9</td>
<td>20</td>
</tr>
<tr>
<td>7</td>
<td>Liquiritin</td>
<td>9</td>
<td>6</td>
<td>8</td>
<td>12</td>
<td>4</td>
<td>13</td>
<td>17</td>
<td>13</td>
</tr>
<tr>
<td>8</td>
<td>Baicalin</td>
<td>17</td>
<td>9</td>
<td>5</td>
<td>14</td>
<td>8</td>
<td>14</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td>9</td>
<td>Isomer of Baicalin</td>
<td>18</td>
<td>10</td>
<td>10</td>
<td>15</td>
<td>19</td>
<td>15</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>10</td>
<td>Oroxylin A-7-O-β-D-glucuronide</td>
<td>16</td>
<td>17</td>
<td>13</td>
<td>16</td>
<td>10</td>
<td>16</td>
<td>8</td>
<td>14</td>
</tr>
<tr>
<td>11</td>
<td>Chrysin-7-O-glucuronide</td>
<td>7</td>
<td>7</td>
<td>9</td>
<td>13</td>
<td>13</td>
<td>9</td>
<td>2</td>
<td>9</td>
</tr>
<tr>
<td>12</td>
<td>Isoliquiritin</td>
<td>10</td>
<td>5</td>
<td>16</td>
<td>20</td>
<td>7</td>
<td>17</td>
<td>13</td>
<td>12</td>
</tr>
<tr>
<td>13</td>
<td>Wogonoside</td>
<td>20</td>
<td>20</td>
<td>11</td>
<td>17</td>
<td>20</td>
<td>5</td>
<td>6</td>
<td>16</td>
</tr>
<tr>
<td>14</td>
<td>Liquiritigenin</td>
<td>19</td>
<td>14</td>
<td>20</td>
<td>11</td>
<td>1</td>
<td>18</td>
<td>19</td>
<td>15</td>
</tr>
<tr>
<td>15</td>
<td>Baicalein</td>
<td>2</td>
<td>11</td>
<td>17</td>
<td>7</td>
<td>6</td>
<td>4</td>
<td>14</td>
<td>3</td>
</tr>
<tr>
<td>16</td>
<td>Isoliquiritigenin</td>
<td>6</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>18</td>
<td>1</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>17</td>
<td>Wogonin</td>
<td>4</td>
<td>12</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>7</td>
<td>16</td>
<td>5</td>
</tr>
<tr>
<td>18</td>
<td>Oroxylin A</td>
<td>5</td>
<td>8</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>8</td>
<td>15</td>
<td>4</td>
</tr>
<tr>
<td>19</td>
<td>Glycyrrhetinic acid</td>
<td>8</td>
<td>1</td>
<td>6</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>18</td>
<td>2</td>
</tr>
<tr>
<td>20</td>
<td>DXMS</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>14</td>
<td>3</td>
<td>20</td>
<td>19</td>
</tr>
</tbody>
</table>
3. Discussions

In order to test the influence of different feature sets on the identification results, we utilized molecular descriptors as control feature set. Molecular descriptors and molecular fingerprints make up full feature set. With these two feature sets, SVM, RF, gcForest and forgeNet are utilized by 3-fold, 5-fold, 10-fold and leave-one-out methods. The AUC and F1 results are depicted in Fig. 9 and Fig. 10, respectively. From the results, it could be seen that full feature set could improve the compound identification accuracy of methods.

4. Conclusions

Network pharmacology has become a frontier and hot spot in the field of traditional Chinese medicine research. This research method can effectively predict the effective components, target and side effects of drugs, and is conducive to the process of modernization of traditional Chinese medicine. In order to construct "disease-gene-target-drug" interaction network more accurately, forest graph embedded deep feed forward network is utilized to infer "disease-compound" network in this paper. According to acute lung injury, two ALI-related target genes (REAL and SATA3) are selected, and the active and inactive compounds of the two corresponding target genes are collected, respectively. Molecular descriptors and molecular fingerprints are utilized to characterize each compound. By leave-one-out, 3-fold, 5-fold and 10-fold cross validation methods, forgeNet has the better performance than SVM, RF and gcForest in terms of $SN$, $SP$, $ACC$, $MCC$, $F1$, AUC and ROC curves. ForgeNet is also utilized to identify 19 compounds in Erhuang decoction (EhD) and Dexamethasone (DXMS) and the results reveal that forgeNet could infer the compounds of disease related more accurately. We also test the influence of different feature sets on the identification results and find the feature set based on molecular descriptors and molecular fingerprints could improve the compound identification accuracy of methods.

5. Methods

5.1 forgeNet

Forest graph-embedded deep feedforward network (forgeNet) is a novel machine learning algorithm, which has been successfully applied to solve classification problem with TCGA RNA-seq data. The flowchart of forgeNet is depicted in Fig. 11. From Fig. 12, it could be seen that this method contains two parts: feature graph construction and deep neural network. Compared with deep learning models, forgeNet solves the dimension problem of biological data and is more robust. The algorithm is described as follows [39].

- Step 1: feature graph construction

Before the labeled training data are input into classifier, the features of the data need to be extracted. In forgeNet, the used forest $\xi$ contains $p$ decision tree (DT). With the labeled training data, the forest is fitted
and \( p \) DT are generated \( (\xi(\theta) = \{ T_1(\theta_1), T_2(\theta_2), \ldots, T_p(\theta_p) \}, \theta_i \) is a parameter). Meanwhile if binary tree is regarded as a special case of directed graph, we can gain the following graph set.

\[
\Phi = \{ G(V_1, E_1), K, G(V_i, E_i), K, G_N(V_p, E_p) \}.
\]

Where \( V_i \) and \( E_i \) represents vertex set and edge set of \( G_i \).

To integrate the directed graph set \( \Phi \), the final aggregated graph can be gained by the following formula.

\[
G = \bigcup_{i=1}^{N} G_i.
\]

2

- Step 2: deep neural network

The feature graph obtained the previous step are embedded into this part. With the processed features graph-embedded deep feedforward networks (GEDFN) is used to train and make the classification for the unknown data [12]. Every layer of GEDFN is introduced as followed.

\[
Z_1 = \sigma(X(W_{in} \Theta G) + b_{in}),
\]

\[
K
\]

\[
Z_{k+1} = \sigma(Z_kW_k + b_k),
\]

\[
K
\]

\[
Z_{out} = \sigma(Z_{out}W_{out} + b_{out}),
\]

\[
y = \text{softmax}(Z_{out}W_{out} + b_{out}).
\]

Where \( X \) is input data, \( Z_k \) is the \( k \) – th hidden layers, \( \Theta \) denotes Hadamard product, \( W_k \) and \( b_k \) are the weights and bias of the \( k \) – th hidden layer, respectively.

5.2 Inference algorithm

(1) Data preparation. Two key target genes: signal transducer and activator of transcription 3 (STAT3), and nuclear transcription factor-\( \kappa \) B/p65 (nuclear factor kappa, B/p65, REAL) were proved to be mainly involved in the key pathways related to acute lung injury (ALI), and loosely related to ALI diseases in the literature [40]. Then, the BindingDB database is searched for the known active compounds of two key target genes [38]. The active ligands are screened with the condition that IC50<5000 nmol·L\(^{-1}\). The collected active compounds are labeled as positive samples. In order to collect the negative samples, 20% of the active ligands are randomly selected and uploaded to DUD-E database (http://dude.docking.org/) to generate the inactive ligands [41]. In order to obtain the molecular descriptors and molecular fingerprints of each ligand, the active and inactive ligands collected are uploaded as the feature vectors.
(2) Model training. According to the collected data, the feature vector of each ligand is used as input for forgeNet. After training phase, the unknown compounds are screened for the target disease.

Declarations

Acknowledgements

This work was supported by the talent project of "Qingtian scholar" of Zaozhuang University, Jiangsu Provincial Natural Science Foundation, China (No. SBK2019040953), Youth Innovation Team of Scientific Research Foundation of the Higher Education Institutions of Shandong Province, China (No. 2019KJM006), the Key Research Program of the Science Foundation of Shandong Province (ZR2020KE001), the PhD research startup foundation of Zaozhuang University (No.2014BS13), and Zaozhuang University Foundation (No. 2015YY02).

Funding

This work was supported by the talent project of "Qingtian scholar" of Zaozhuang University, Jiangsu Provincial Natural Science Foundation, China (No. SBK2019040953), Youth Innovation Team of Scientific Research Foundation of the Higher Education Institutions of Shandong Province, China (No. 2019KJM006), the Key Research Program of the Science Foundation of Shandong Province (ZR2020KE001), the PhD research startup foundation of Zaozhuang University (No.2014BS13), and Zaozhuang University Foundation (No. 2015YY02).

Availability of data and materials

The data that support the findings of this study are available on request from the corresponding author.

Declarations Ethics approval and consent to participate

Not applicable.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

All authors have no conflict of interest to declare.

References


31. Liu Yi, Li Fei, Jia Yuejin, Hao Shifei, Exploration of Mechanisms of Danzhixiaoyao Powder in Treatment of Insomnia Disorder Based on Network Pharmacology and Molecular Docking, Clinical


Figures

Figure 1

ROC Performances of four methods with Data1 by 3-fold cross validation method.
Figure 2

ROC Performances of four methods with Data1 by 5-fold cross validation method.
Figure 3

ROC Performances of four methods with Data1 by 10-fold cross validation method.
Figure 4

ROC Performances of four methods with Data1 by leave-one-out method.
Figure 5

ROC Performances of four methods with Data2 by 3-fold cross validation method.
Figure 6

ROC Performances of four methods with Data2 by 5-fold cross validation method.
Figure 7

ROC Performances of four methods with Data2 by 10-fold cross validation method.
Figure 8

ROC Performances of four methods with Data2 by leave-one-out method.
Figure 9

AUC performances of four methods by 3-fold (a), 5-fold (b), 10-fold (c) and leave-one-out (d) methods with full feature set (blue) and control feature set (red).
Figure 10

F1 performances of four methods by leave-one-out (a), 3-fold (b), 5-fold (c), and 10-fold (d) methods with full feature set (blue) and control feature set (red).
Figure 11

The flowchart of forgeNet algorithm.
Figure 12

The flowchart of drug components-disease network related to acute lung injury inference.