Machine Learning-based Classification of Colorectal Cancer Stage Diagnosis Using Recurrent Neural Network Model Algorithm

Rifaldy Fajar, Nana Indri, and Titik Wulandari

Computational Biology Laboratory,
Yogyakarta State University, Indonesia
With early diagnosis and treatment, the prognosis of early stage colorectal cancer patients is optimistic, and the middle and long-term survival rate is usually high.
One way to detect colorectal cancer is by using colonoscopy. Once colorectal cancer is detected, a classification is performed to determine the stage of cancer.

In this study, the Recurrent Neural Network (RNN) model was used for the classification of colorectal cancer stages. This study aims to explain the procedure and the accuracy of Elman tissue RNN modeling in the classification of colorectal cancer stages from colonoscopy photographs.
Method

Data Sources

The data used are secondary data in the form of colonoscopy photos of the colon or rectum. The colonoscopy results have been classified based on the stage. 92 photo data from colonoscopy, this data is obtained from health documentation available at:

1. Vision and Image Analysis Group, Cornell University (Murra-Saca)
2. Endoskopie Atlas, Albertinen-Krankenhaus

The composition of the data used is 80% (72 data) as training data and 20% (18 data) as testing data. The division of the data is not unanimous because the total data is 92 data. The composition of this data sharing is based on the MSE value and the accuracy between training data and testing data. The composition of this data sharing results in smaller MSE values and accuracy.
The colonoscopy data is processed by extracting the image. The image extraction process can be done with the Gray Level Cooccurrence Matrix (GLCM). The statistical parameters were collected using MATLAB with an Image Processing Toolbox and a prototype system designed using a Graphical User Interface (GUI) to make it easier.
Method

The input variable consists of 14 statistical parameters from the extraction of colonoscopy images, namely:
1. Energy (X1)
2. contrast (X2)
3. correlation (X3)
4. sum of squares (X4)
5. inverse difference moment (X5)
6. sum average (X6)
7. sum variance (X7)
8. sum entropy (X8)
9. Entropy (X9)
10. difference variance (X10)
11. difference entropy (X11)
12. maximum probability (X12)
13. Homogeneity (X13)
14. Dissimilarity (X14)

Based on the results of image extraction, it is obtained two statistical parameters that produce the same value, namely contrast and difference variance. This causes one statistical parameter to be eliminated, namely Difference Variance (X10) so that the input variable becomes 13 variables.

Classification Criteria

The output variables are numerical data, and the expected output is a number 1 for normal (not affected by colorectal cancer), 2 for stage I colorectal cancer, 3 for stage II colorectal cancer, 4 for stage III colorectal cancer, and 5 for stage IV colorectal cancer.

If the output \( (y) \) is \( 0.5 \leq y < 1.5 \) then it is considered 1, so the classification result is normal.

If the output \( (y) \) is \( 1.5 \leq y < 2.5 \) then it is considered 2, so the classification result is stage I.

If the output \( (y) \) is \( 2.5 \leq y < 3.5 \) then it is considered 3, so the classification result is stage II.

If the output \( (y) \) is \( 3.5 \leq y < 4.5 \) then it is considered 4, so the classification result is stage III.

If the output \( (y) \) is \( 4.5 \leq y < 5.5 \) then it is considered 5, so the classification result is stage IV.
Results

The best model is obtained with 9 inputs $x_1, x_2, x_3, x_5, x_7, x_9, x_{12}, x_{13}, x_{14}$ and 1 neuron in the hidden layer with a bipolar sigmoid activation $x_1$ function (tansig) in the hidden layer, and a linear (purelin) function in the output layer.

The 74 training data shows 53 true positive data (TP), 4 false positive data (FP), 15 true negative data (TN) and 2 false negative data (FN).

So that we get 96% sensitivity level, 79% level of specificity, 92% level of accuracy of truth from the RNN model. Meanwhile, the 18 testing data showed 100% sensitivity level, 80% specificity level, 94% accuracy level.
Results

23 out of 72 images

Some colonoscopy images of training data (Real Image and Grayscale)
### Results

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Normal</td>
<td><img src="image1" alt="Real Image" /></td>
<td><img src="image2" alt="Grayscale" /></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Normal</td>
<td><img src="image3" alt="Real Image" /></td>
<td><img src="image4" alt="Grayscale" /></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Normal</td>
<td><img src="image5" alt="Real Image" /></td>
<td><img src="image6" alt="Grayscale" /></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Normal</td>
<td><img src="image7" alt="Real Image" /></td>
<td><img src="image8" alt="Grayscale" /></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Normal</td>
<td><img src="image9" alt="Real Image" /></td>
<td><img src="image10" alt="Grayscale" /></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Stadium1</td>
<td><img src="image11" alt="Real Image" /></td>
<td><img src="image12" alt="Grayscale" /></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Stadium1</td>
<td><img src="image13" alt="Real Image" /></td>
<td><img src="image14" alt="Grayscale" /></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Stadium1</td>
<td><img src="image15" alt="Real Image" /></td>
<td><img src="image16" alt="Grayscale" /></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Stadium1</td>
<td><img src="image17" alt="Real Image" /></td>
<td><img src="image18" alt="Grayscale" /></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Stadium2</td>
<td><img src="image19" alt="Real Image" /></td>
<td><img src="image20" alt="Grayscale" /></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Stadium2</td>
<td><img src="image21" alt="Real Image" /></td>
<td><img src="image22" alt="Grayscale" /></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Stadium2</td>
<td><img src="image23" alt="Real Image" /></td>
<td><img src="image24" alt="Grayscale" /></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Stadium2</td>
<td><img src="image25" alt="Real Image" /></td>
<td><img src="image26" alt="Grayscale" /></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Stadium3</td>
<td><img src="image27" alt="Real Image" /></td>
<td><img src="image28" alt="Grayscale" /></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Stadium3</td>
<td><img src="image29" alt="Real Image" /></td>
<td><img src="image30" alt="Grayscale" /></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Stadium4</td>
<td><img src="image31" alt="Real Image" /></td>
<td><img src="image32" alt="Grayscale" /></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Stadium4</td>
<td><img src="image33" alt="Real Image" /></td>
<td><img src="image34" alt="Grayscale" /></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Stadium4</td>
<td><img src="image35" alt="Real Image" /></td>
<td><img src="image36" alt="Grayscale" /></td>
<td></td>
</tr>
</tbody>
</table>

18 out of 18 images [All images]

All colonoscopy images of testing data (Real Image and Grayscale)
Conclusion

The machine learning classification based on the RNN model algorithm can be said to be good enough to determine the stage level of colorectal cancer.