Introduction:

Objectives:

Methods:

Workflow:

Features:

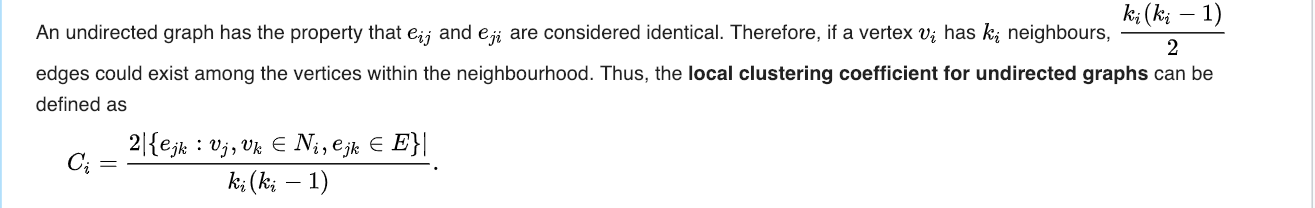
**Topological Features:**

1. Average Shortest Path:

Step 1: Find the geodesic distance from a non-disease node to each disease node

Step 2: Take the average of all distances as the Average Shortest Path of a non-disease node to all disease nodes

1. Local Clustering Coefficient:

Implemented according to the formula

The cluster E is the disease cluster, and N is the non-disease nodes cluster

1. Degree Centrality

Implemented by NetworkX library

Rationale: Higher, more likely to be involved in a more important functional module

1. Closeness Centrality

Implemented by NetworkX library

Rationale: Higher, more functionally important as needs to be communicated quickly.

1. Betweeness Centrality

Implemented by NetworkX library

Rationale: An important node will lie on a higher proportion of the paths.

1. Eigenvector Centrality

Implemented by NetworkX library

Rationale: the influence of a node in a network

1. Percolation centrality

Implemented by NetworkX library. This method has problems; thus the library method couldn’t successfully be run.

Rationale: importance of a node in purely topological terms, despite the network dynamic

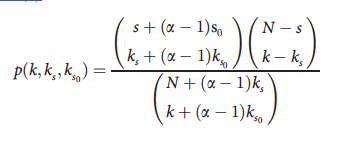
1. Pagerank

Implemented by NetworkX library.

Rationale: the notion of how central a node is in a network relative to a particular node

1. Connectivity Significance

Implemented using the formula given in the DIAMOnD paper



1. Articulation Point

Implemented using the algorithm given in this website (<https://www.geeksforgeeks.org/articulation-points-or-cut-vertices-in-a-graph/> ), the code has been slightly modified.

1. Modularity

Step 1: NetworkX library computes the modularity matrix

Step 2: Summing over all modularity from a node (row, in Gc.nodes()) to a disease node (column), exclude the node itself.

Step 3: Take the average for each node ( in Gc.nodes()).

**Sequence based features:**

**Functional Features:**

**Disease Module Identification**

**Dataset (After Data Cleaning, Data Balancing and Stratified Split)**

|  |  |  |
| --- | --- | --- |
| Train set | | |
| Disease Node | Non Disease Node (Essential and Non-essential) | Total |
| 987 | 3033 | 4020 |
| Test set | | |
| Disease Node | Non Disease Node (Essential and Non-essential) | Total |
| 454 | 1296 | 1723 |

**Dataset 1**

|  |  |  |
| --- | --- | --- |
| Train set | | |
| Disease Node | Non Disease Node (Essential and Non-essential) | Total |
| 1009 | 1008 | 2017 |
| Test set | | |
| Disease Node | Non Disease Node (Essential and Non-essential) | Total |
| 432 | 433 | 865 |

**Dataset 2**

|  |  |  |
| --- | --- | --- |
| Train set | | |
| Disease Node | Non Disease Node (Essential and Non-essential) | Total |
| 1009 | 1008 | 2017 |
| Test set | | |
| Disease Node | Non Disease Node (Essential and Non-essential) | Total |
| 432 | 433 | 865 |

**Dataset 3**

|  |  |  |
| --- | --- | --- |
| Train set | | |
| Disease Node | Non Disease Node (Essential and Non-essential) | Total |
| 1008 | 994 | 2002 |
| Test set | | |
| Disease Node | Non Disease Node (Essential and Non-essential) | Total |
| 433 | 426 | 859 |

**Experiment1: K means clustering with Principle Component Analysis (on the test set)**

* Dataset 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Condition | Accuracy | Precision | Recall | F1 |
| Cluster 0 is considered as target | 0.27630058 | 0.00510204 | 0.00231481 | 0.00318471 |
| Cluster 0 is considered as non target | 0.72369942 | 0.64424514 | 0.99768519 | 0.78292461 |
| Cluster 1 is considered as target | 0.96069364 | 0.93073593 | 0.99537037 | 0.96196868 |
| Cluster 1 is considered as non target | 0.03930636 | 0.00496278 | 0.00462963 | 0.00479042 |
| Cluster 2 is considered as target | 0.26358382 | 0.00483092 | 0.00231481 | 0.00312989 |
| Cluster 2 is considered as non target | 0.73641618 | 0.6550152 | 0.99768519 | 0.79082569 |

* Dataset 2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Condition | Accuracy | Precision | Recall | F1 |
| Cluster 0 is considered as target | 0.30635838 | 0.01724138 | 0.00694444 | 0.00990099 |
| Cluster 0 is considered as non target | 0.69364162 | 0.62083936 | 0.99305556 | 0.76402493 |
| Cluster 1 is considered as target | 0.95260116 | 0.91862955 | 0.99305556 | 0.95439377 |
| Cluster 1 is considered as non target | 0.04739884 | 0.00753769 | 0.00694444 | 0.00722892 |
| Cluster 2 is considered as target | 0.2416185 | 0 | 0 | 0 |
| Cluster 2 is considered as non target | 0.7583815 | 0.67394696 | 1 | 0.80521901 |

* Dataset 3

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Condition | Accuracy | Precision | Recall | F1 |
| Cluster 0 is considered as target | 0.26076834 | 0.00490196 | 0.00230947 | 0.00313972 |
| Cluster 0 is considered as non target | 0.73923166 | 0.65954198 | 0.99769053 | 0.79411765 |
| Cluster 1 is considered as target | 0.9685681 | 0.94713656 | 0.99307159 | 0.96956032 |
| Cluster 1 is considered as non target | 0.0314319 | 0.00740741 | 0.00692841 | 0.0071599 |
| Cluster 2 is considered as target | 0.26658906 | 0.00995025 | 0.00461894 | 0.00630915 |
| Cluster 2 is considered as non target | 0.73341094 | 0.6550152 | 0.99538106 | 0.79010082 |

**Experiment2: Support Vector Machine with RBF Kernel**

* Dataset 1

Cross Validation Score

|  |  |  |
| --- | --- | --- |
| CV1 | CV2 | CV3 |
| 1 | 1 | 1 |

Test Set Score

|  |  |  |  |
| --- | --- | --- | --- |
| Accuracy | Precision | Recall | F1 Score |
| 0.99768786 | 0.99768519 | 0.99768519 | 0.99768519 |

* Dataset 2

Cross Validation Score

|  |  |  |
| --- | --- | --- |
| CV1 | CV2 | CV3 |
| 0.99405646 | 0.99553571 | 0.99702381 |

Test Set Score

|  |  |  |  |
| --- | --- | --- | --- |
| Accuracy | Precision | Recall | F1 Score |
| 0.99653179 | 1 | 0.99305556 | 0.99651568 |

* Dataset 3

Cross Validation Score

|  |  |  |
| --- | --- | --- |
| CV1 | CV2 | CV3 |
| 0.99850299 | 0.994003 | 0.994003 |

Test Set Score

|  |  |  |  |
| --- | --- | --- | --- |
| Accuracy | Precision | Recall | F1 Score |
| 1 | 1 | 1 | 1 |

**Experiment3: Random Forest**

* Dataset 1

Cross Validation Score

|  |  |  |
| --- | --- | --- |
| CV1 | CV2 | CV3 |
| 1 | 1 | 0.9985119 |

Test Set Score

|  |  |  |  |
| --- | --- | --- | --- |
| Accuracy | Precision | Recall | F1 Score |
| 1 | 1 | 1 | 1 |

* Dataset 2

Cross Validation Score

|  |  |  |
| --- | --- | --- |
| CV1 | CV2 | CV3 |
| 0.99851412 | 0.99553571 | 1 |

Test Set Score

|  |  |  |  |
| --- | --- | --- | --- |
| Accuracy | Precision | Recall | F1 Score |
| 0.99653179 | 1 | 0.99305556 | 0.99651568 |

* Dataset 3

Cross Validation Score

|  |  |  |
| --- | --- | --- |
| CV1 | CV2 | CV3 |
| 0.99850299 | 0.99550225 | 0.99550225 |

Test Set Score

|  |  |  |  |
| --- | --- | --- | --- |
| Accuracy | Precision | Recall | F1 Score |
| 1 | 1 | 1 | 1 |

**Experiment4: Deep Learning**

* Dataset 1

Cross Validation Score

|  |  |  |
| --- | --- | --- |
| CV1 | CV2 | CV3 |
| 0.85714286 | 1 | 0.99107143 |

Test Set Score

|  |  |  |  |
| --- | --- | --- | --- |
| Accuracy | Precision | Recall | F1 Score |
|  |  |  |  |

* Dataset 2

Cross Validation Score

|  |  |  |
| --- | --- | --- |
| CV1 | CV2 | CV3 |
| 0.97321429 | 0.98511905 | 0.99107143 |

Test Set Score

|  |  |  |  |
| --- | --- | --- | --- |
| Accuracy | Precision | Recall | F1 Score |
|  |  |  |  |

* **Dataset 3**

Cross Validation Score

|  |  |  |
| --- | --- | --- |
| CV1 | CV2 | CV3 |
| 0.74850299 | 1 | 0.93403298 |

Test Set Score

|  |  |  |  |
| --- | --- | --- | --- |
| Accuracy | Precision | Recall | F1 Score |
|  |  |  |  |

**Experiment5: Conventional module detection**