Introduction:

Objectives:

Methods:

Workflow:

Features:

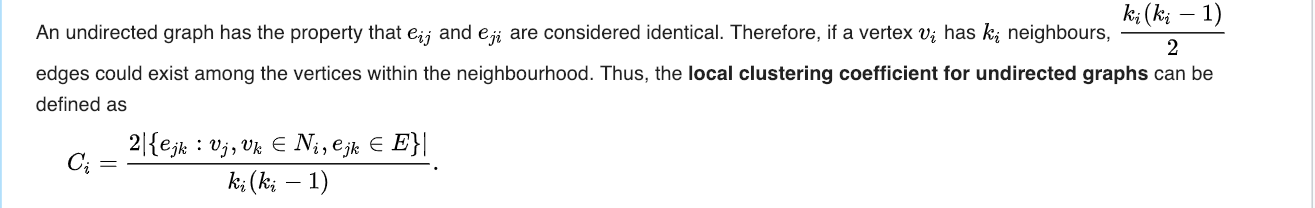
**Topological Features:**

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

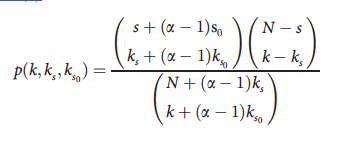
1. Pagerank

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)**

|  |  |  |
| --- | --- | --- |
| Disease Node | Non Disease Node (Essential and Non-essential) | Total |
| 1441 | 4329 | 5770 |

**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 whole dataset)**

* Dataset 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Condition | Accuracy | Precision | Recall | F1 |
| Cluster 0 is considered as disease | 0.45107564 | 0.41198502 | 0.22900763 | 0.29438002 |
| Cluster 0 is considered as non-disease | 0.54892436 | 0.53387794 | 0.77099237 | 0.63089154 |
| Cluster 1 is considered as disease | 0.52602359 | 0.52856055 | 0.48160999 | 0.50399419 |
| Cluster 1 is considered as non-disease | 0.47397641 | 0.47609943 | 0.51839001 | 0.49634551 |
| Cluster 2 is considered as disease | 0.52290076 | 0.54296875 | 0.28938237 | 0.3775464 |
| Cluster 2 is considered as non-disease | 0.47709924 | 0.48438978 | 0.71061763 | 0.57609001 |

Conclusion: Disease set: cluster 0; Non-disease set: cluster 1 and cluster 2

* Dataset 2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Condition | Accuracy | Precision | Recall | F1 |
| Cluster 0 is considered as disease | 0.45697432 | 0.42091837 | 0.22900763 | 0.29662921 |
| Cluster 0 is considered as non-disease | 0.54302568 | 0.52955195 | 0.77099237 | 0.62786098 |
| Cluster 1 is considered as disease | 0.51804303 | 0.53412073 | 0.28244275 | 0.36949614 |
| Cluster 1 is considered as non-disease | 0.48195697 | 0.48773585 | 0.71755725 | 0.58073575 |
| Cluster 2 is considered as disease | 0.52498265 | 0.52694611 | 0.48854962 | 0.50702197 |
| Cluster 2 is considered as non-disease | 0.47501735 | 0.4767141 | 0.51145038 | 0.49347171 |

Conclusion: Disease set: cluster 0; Non-disease set: cluster 1 and cluster 2

* Dataset 3

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Condition | Accuracy | Precision | Recall | F1 |
| Cluster 0 is considered as disease | 0.52079692 | 0.54475703 | 0.29562804 | 0.38326586 |
| Cluster 0 is considered as non-disease | 0.47920308 | 0.48821549 | 0.70437196 | 0.57670455 |
| Cluster 1 is considered as disease | 0.44669696 | 0.40482574 | 0.20957668 | 0.27617741 |
| Cluster 1 is considered as non-disease | 0.55330304 | 0.53853428 | 0.79042332 | 0.64060742 |
| Cluster 2 is considered as disease | 0.52883607 | 0.53488372 | 0.49479528 | 0.51405912 |
| Cluster 2 is considered as non-disease | 0.47116393 | 0.47643979 | 0.50520472 | 0.49040081 |

Conclusion: Disease set: cluster 0; Non-disease set: cluster 1 and cluster 2

* Jaccard Similarity between the disease set among three different dataset

|  |  |  |  |
| --- | --- | --- | --- |
|  | Data set 1 | Data set 2 | Data set 3 |
| Data set 1 | 1 | 0.19558359621451105 | 0.00505369551484523 |
| Data set 2 | 0.19558359621451105 | 1 | 0.011494252873563218 |
| Data set 3 | 0.00505369551484523 | 0.011494252873563218 | 1 |

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

* Dataset 1

Cross Validation Score

|  |  |
| --- | --- |
| CV1 | 0.62376238 |
| CV2 | 0.60891089 |
| CV3 | 0.61881188 |
| CV4 | 0.54455446 |
| CV5 | 0.58910891 |
| CV6 | 0.59405941 |
| CV7 | 0.58415842 |
| CV8 | 0.58415842 |
| CV9 | 0.51741294 |
| CV10 | 0.48 |

Test Set Score

|  |  |  |  |
| --- | --- | --- | --- |
| Accuracy | Precision | Recall | F1 Score |
| 0.55491329 | 0.55119826 | 0.58564815 | 0.56790123 |

* Dataset 2

Cross Validation Score

|  |  |
| --- | --- |
| CV1 | 0.55445545 |
| CV2 | 0.53960396 |
| CV3 | 0.57920792 |
| CV4 | 0.56435644 |
| CV5 | 0.56930693 |
| CV6 | 0.55940594 |
| CV7 | 0.64356436 |
| CV8 | 0.55940594 |
| CV9 | 0.60696517 |
| CV10 | 0.55 |

Test Set Score

|  |  |  |  |
| --- | --- | --- | --- |
| Accuracy | Precision | Recall | F1 Score |
| 0.59306358 | 0.5952381 | 0.5787037 | 0.58685446 |

* Dataset 3

Cross Validation Score

|  |  |
| --- | --- |
| CV1 | 0.51741294 |
| CV2 | 0.53233831 |
| CV3 | 0.59701493 |
| CV4 | 0.62686567 |
| CV5 | 0.66 |
| CV6 | 0.66 |
| CV7 | 0.61 |
| CV8 | 0.55 |
| CV9 | 0.6281407 |
| CV10 | 0.5678392 |

Test Set Score

|  |  |  |  |
| --- | --- | --- | --- |
| Accuracy | Precision | Recall | F1 Score |
| 0.56810244 | 0.562 | 0.64896074 | 0.60235798 |

**Experiment3: Random Forest**

* Dataset 1

Cross Validation Score

|  |  |
| --- | --- |
| CV1 | 0.71287129 |
| CV2 | 0.72277228 |
| CV3 | 0.70792079 |
| CV4 | 0.73267327 |
| CV5 | 0.68316832 |
| CV6 | 0.66831683 |
| CV7 | 0.72277228 |
| CV8 | 0.71287129 |
| CV9 | 0.68159204 |
| CV10 | 0.64 |

Test Set Score

|  |  |  |  |
| --- | --- | --- | --- |
| Accuracy | Precision | Recall | F1 Score |
| 0.68786127 | 0.68786127 | 0.68287037 | 0.68604651 |

* Dataset 2

Cross Validation Score

|  |  |
| --- | --- |
| CV1 | 0.62871287 |
| CV2 | 0.62376238 |
| CV3 | 0.67821782 |
| CV4 | 0.65346535 |
| CV5 | 0.65841584 |
| CV6 | 0.62871287 |
| CV7 | 0.65841584 |
| CV8 | 0.65841584 |
| CV9 | 0.70646766 |
| CV10 | 0.66 |

Test Set Score

|  |  |  |  |
| --- | --- | --- | --- |
| Accuracy | Precision | Recall | F1 Score |
| 0.69710983 | 0.69767442 | 0.69444444 | 0.69605568 |

* Dataset 3

Cross Validation Score

|  |  |
| --- | --- |
| CV1 | 0.6119403 |
| CV2 | 0.71641791 |
| CV3 | 0.69154229 |
| CV4 | 0.68159204 |
| CV5 | 0.715 |
| CV6 | 0.75 |
| CV7 | 0.7 |
| CV8 | 0.625 |
| CV9 | 0.69346734 |
| CV10 | 0.66834171 |

Test Set Score

|  |  |  |  |
| --- | --- | --- | --- |
| Accuracy | Precision | Recall | F1 Score |
| 0.67520373 | 0.6673913 | 0.70900693 | 0.68756999 |

**Experiment4: Deep Learning**

* Dataset 1

Cross Validation Score

|  |  |
| --- | --- |
| CV1 |  |
| CV2 |  |
| CV3 |  |
| CV4 |  |
| CV5 |  |
| CV6 |  |
| CV7 |  |
| CV8 |  |
| CV9 |  |
| CV10 |  |

Test Set Score

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

* Dataset 2

Cross Validation Score

|  |  |
| --- | --- |
| CV1 |  |
| CV2 |  |
| CV3 |  |
| CV4 |  |
| CV5 |  |
| CV6 |  |
| CV7 |  |
| CV8 |  |
| CV9 |  |
| CV10 |  |

Test Set Score

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

* **Dataset 3**

Cross Validation Score

|  |  |
| --- | --- |
| CV1 |  |
| CV2 |  |
| CV3 |  |
| CV4 |  |
| CV5 |  |
| CV6 |  |
| CV7 |  |
| CV8 |  |
| CV9 |  |
| CV10 |  |

Test Set Score

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

**Experiment5: External Validation**