

# A Network Learning Method for Functional Disability Prediction from Health Data

Riccardo Dondi<sup>1</sup><sup>a</sup> and Mehdi Hosseinzadeh<sup>1,2</sup><sup>b</sup>

<sup>1</sup>Università degli Studi di Bergamo, Bergamo, Italy

<sup>2</sup>University of Calabria, Rende(CS), Italy

**Keywords:** Network Analysis, Disability Classification, Learning Algorithms, Healthcare Analytics, Graph Data Mining.

**Abstract:** This contribution proposes a novel network analysis model with the goal of predicting a classification of individuals as either 'disabled' or 'not-disabled', using a dataset from the Health and Retirement Study (HRS). Our approach is based on selecting features that span health indicators and socioeconomic factors due to their pivotal roles in identifying disability. Considering the selected features, our approach computes similarities between individuals and uses this similarity to predict disability. We present a preliminary experimental evaluation of our method on the HRS dataset, where it shows an enhanced average accuracy of 62.48%.

## 1 INTRODUCTION

A relevant problem for supporting elderly individuals is the prediction of their health status. In this context, it is extremely valuable to predict the risk of functionally disability, in order to provide the needed support (Stuck et al., 1999).


Current studies demonstrate the advancements in the use of knowledge graphs and network analysis in the fields of biology and healthcare (Hosseinzadeh, 2020; Hosseinzadeh et al., 2022) and (Pham et al., 2018; Tao et al., 2020; Wang et al., 2020; Pham et al., 2022; Cui et al., 2023). In this context, the development of prediction models based on graphs in healthcare is essential for improving disease diagnosis and reducing human error. In particular, (Wang et al., 2020) developed a predictive model that classifies individuals according to their disability risk, using a network to represent disease progression. (Tao et al., 2020) introduced a novel classification model that uses a heterogeneous knowledge graph for conceptualizing medical domain knowledge. The developed model was used to forecast possible health risks for patients using data from the National Health and Nutrition Examination Survey (NHANES). (Cui et al., 2023) provided a comprehensive review of knowledge graph applications in healthcare, highlighting the instruments, applications, and possibilities for


improved understanding and prediction of complex medical scenarios.

Our contribution aims to build a prediction method inspired by approaches for classifying individuals based on their risk of becoming disabled. Our approach proposes a novel network analysis model based on the features presented in a dataset from the Health and Retirement Study (HRS)<sup>1</sup> (Health and Retirement Study, 2008). We select some features that span health indicators and socioeconomic factors due to their pivotal roles in identifying disability. Each individual is then represented as a vector on the selected features and similarity between two individuals is evaluated by computing a function of the difference in the values of the features. A user is then assigned to the category ('disabled', meaning high-risk of becoming disabled, or 'non-disabled', meaning low-risk of becoming disabled) based on the average similarity with each single group (disabled individuals and non disabled individuals).

We present some preliminary experimental evaluation of our method on the HRS dataset. We select 10 samples of 100 individuals extracted randomly from the HRS dataset and on each of this sample we evaluate the performance of our method. The method shows a moderate accuracy in the classification (average accuracy of 62.48%).

The remainder of the paper is organized as follows. In section 2, we start by introducing some defi-

<sup>a</sup> <https://orcid.org/0000-0002-6124-2965>

<sup>b</sup> <https://orcid.org/0000-0003-3275-6286>

<sup>1</sup><https://hrs.isr.umich.edu>.

nitions and by providing the formal definition by formally introduces the research problem. In Section 3, we present the computational approach used to address the research problem, including the construction and application of the bipartite graph model. In Section 4, we present the results from the experimental analysis, discussing the implications and insights gained from applying our methodology to the HRS dataset. Finally, In Section 5, we conclude the main outcomes with some future directions.

## 2 DEFINITIONS AND RESEARCH PROBLEM

In this section, we define the main concepts needed for our methodology, mainly graph theory and network analysis, and we present the formal research problem our study addresses.

All the graphs we consider in this paper are undirected. A graph  $G$  is defined as a pair  $G = (V, E)$ , where  $V$  is a set of nodes and  $E$  is a set of edges. Each edge  $e \in E$  is an unordered pair  $\{v, w\}$ , indicating a connection between nodes  $v$  and  $w$  in  $V$  (Bondy and Murty, 2008). We mainly consider bipartite graph, defined in the following.

**Definition 1.** A graph is bipartite if there exist two disjoint sets  $X \subseteq V$  and  $Y \subseteq V$  such that  $X \uplus Y = V^2$ , and every edge in  $E$  links a node of  $X$  and a node of  $Y$ .

We now provide the definition of the neighborhood of a node, which is a relevant concept needed for describing our method.

**Definition 2.** Given a graph  $G = (V, E)$  and a node  $v \in V$ , the neighborhood of  $v$  in  $G$  is defined as  $N_G(v) = \{u \in V \mid \{u, v\} \in E\}$ .

We now introduce a specific bipartite weighted graph we consider to represent the relations between features and individuals, called *Feature-Individual Classification Network* (FICN).

**Definition 3.** The *Feature-Individual Classification Network* is a bipartite weighted graph denoted as  $G = (V, E, W_E, W_F)$  where:

- $V$  is the set of nodes, partitioned into two disjoint subsets  $V_I$  and  $V_F$ . The subset  $V_I$  represents individuals, and  $V_F$  represents features.
- $E \subseteq V_I \times V_F$  is the set of edges, each connecting an individual node in  $V_I$  and a feature node in  $V_F$ .
- $W_E : E \rightarrow \mathbb{R}^+$  is a weight function for the edges, where each weight represents the value of an individual for a specific feature (these weights are

<sup>2</sup>We recall that  $\uplus$  denotes the disjoint union of sets.

derived from the input data, their computation is described later).

- $W_F : V_F \rightarrow \mathbb{R}^+$  is a weight function for the feature nodes, assigning a weight to each feature; this weight represents the relevance of the specific feature for classification. Unlike  $W_E$ ,  $W_F$  is not obtained from the input data but it is computed using an optimization technique (that we will describe in Section 3).

Given an edge  $uv \in E$ ,  $w_E(uv)$  denotes the edge weight of  $uv$ . Given a node  $u \in V_F$ ,  $w_F(u)$  denotes the feature weight of node  $u$ .

Note that  $G = (V, E, W_E, W_F)$  is not a complete bipartite graph as some edges may not be defined, reflecting possible missing data of individuals.

The classification of an individual is based on the similarity value between the node related to the individual, and the nodes of the Feature-Individual Classification Network, as defined in the following.

**Definition 4.** Let  $G = (V, E, W_E, W_F)$  be a *Feature-Individual Classification Network*, where  $V = V_I \uplus V_F$ . Consider a candidate node  $c$  (note that  $c \notin V$ ) such that  $c$  is connected with a subset  $F_c$  of feature nodes (hence  $F_c \subseteq V_F$ ). The similarity measure  $\sigma(c, G)$  of  $c$  with respect to  $G$  is defined as:

$$\sigma(c, G) = \frac{\sum_{f \in F_c} w_I(f) z_{cf}}{\sum_{f \in F_c} w_I(f)},$$

where  $w_I(f)$  is the weight function of individual  $I$  for feature  $f$ , and for each  $f \in N_G(c)$ ,  $z_{cf}$  is the z-score<sup>3</sup> of  $w_E(cf)$  in the following set:

$$\{w_E(uf) : u \in N_G(f)\}.$$

Note that the similarity measure  $\sigma(c, G)$  is based on the weight of the features that are computed by the optimization technique described in Section 3.

### 2.1 Research Problem

Next we describe our problem. Given two disjoint sets of individuals ('disable' and 'non-disable'), we define a Feature-Individual Classification Network for each of these sets.

$$G_1 = (V_{I_1} \uplus V_F, E_1, W_{E_1}, W_{F_1})$$

represents the graph consisting of the set  $V_{I_1}$  of individuals identified as 'disabled'.

$$G_2 = (V_{I_2} \uplus V_F, E_2, W_{E_2}, W_{F_2})$$

<sup>3</sup>The z-score for feature  $f$  is computed as  $z_{cf} = \frac{v_{cf} - \mu_f}{\sigma_f}$ , where  $v_{cf}$  is the value of feature  $f$  for the candidate node  $c$ ,  $\mu_f$  is the mean value of feature  $f$  across all the neighbor nodes  $N_G(f)$  in  $G$  that have  $f$ , and  $\sigma_f$  is the standard deviation of  $f$  among the same nodes.

represents the set  $V_2$  of ‘non-disabled’ individuals.

We introduce now the main research problem we consider in this paper, which aims to compute the feature weights in order to optimize the classification.

**Problem 1. Weight Feature Optimization Problem.**

**Input:** Two Feature-Individual Classification Networks  $G_1 = (V_1 \uplus V_F, E_1, W_{E_1}, W_F)$  and  $G_2 = (V_2 \uplus V_F, E_2, W_{E_2}, W_F)$ .

**Output:** Compute  $W_F : V_F \rightarrow \mathbb{R}^+$  for the feature nodes so that the classification in ‘disabled’ or ‘non-disabled’ individuals is optimized.

Assume that the feature weights are known. For each individual  $i_{\text{class}}$  to be classified, we compute the similarity  $\sigma(i_{\text{class}}, G_i)$ , with  $i \in \{1, 2\}$ . After calculating these similarity scores, the overall classification of  $i_{\text{class}}$  as either ‘disabled’ or ‘non-disabled’ is obtained by aggregating these scores:

$$\text{Classification}(i_{\text{class}}) = \arg \max_{G \in \{G_1, G_2\}} \sigma(i_{\text{class}}, G).$$

This aggregate score assigns  $i_{\text{class}}$  to the group with the highest computed similarity score. So, in order to apply the classification, we need to compute the feature weights.

The Weight Feature Optimization Problem is solved by considering a training set of individuals for which we already know the classification and then compute the value of the weights  $W_F$  in order to maximize the correct classification. Formally, we consider the following problem.

**Problem 2. Training Weight Feature Optimization Problem.**

**Input:** Two Feature-Individual Classification Networks  $G_1 = (V_1 \uplus V_F, E_1, W_{E_1}, W_F)$  and  $G_2 = (V_2 \uplus V_F, E_2, W_{E_2}, W_F)$ ; two sets  $X_1, X_2$  of candidate nodes that are classified as disable and non-disable, respectively.

**Output:** Compute  $W_F : V_F \rightarrow \mathbb{R}^+$  so that the number of individuals of  $X_1 \uplus X_2$  correctly classified is maximized.

### 3 METHODOLOGY

In the preliminary phase of our study, we define our classification criteria based on the established guidelines from (Li et al., 2017; Rossetti and Cazabet, 2018). Specifically, an individual is considered ‘disabled’ when encounters two or more difficulties in any of the six identified Activities of Daily Living (ADL). In order to address the classification problem, we structure our data into two disjoint sets, i.e. a training set and a test set.

The training set consists of distinct subsets for different phases of the model development process. The first subset of the training set consists of (1) 250 individuals randomly selected from the ‘disabled’ individuals and used to build the Feature-Individual Classification Network  $G_1$  representing ‘disabled’ individuals, (2) 250 individuals randomly selected from the ‘disabled’ individuals and used to build  $G_2$  representing ‘non-disabled’ individuals. The second subset consists of 100 ‘disabled’ individuals, and 100 ‘non-disabled’ individuals used for the weight optimization phase of our model. Note that the first and second subsets are disjoint.

The test set is a subset consisting of individuals whose disability status is also known; it is not utilized to compute feature weights, but for assessing the accuracy of our model. This set will be described in Section 4.

In order to solve the Training Weight Feature Optimization Problem, we implemented an optimization technique. This process starts with uniform initial weights for each feature. Then we adopt a greedy algorithm to incrementally adjust these weights, one at a time. This process is mathematically formulated as follows:

**Initial Setup:** The optimization starts with uniform initial weights for each feature:  $W_F(f) = 1$  for all  $f \in V_F$ , where recall that  $V_F$  is the set of all feature nodes.

**Greedy Algorithm for Weight Adjustment:** We adopt a greedy algorithm to incrementally adjust these weights, where each iteration focuses on changing the value of a single feature weight. The adjustment process is mathematically formulated as follows:

$$W_F(f) \leftarrow W_F(f) + \Delta w_f,$$

where  $\Delta w_f$  is the change in weight for feature  $f$ . After applying this change, we evaluate its effectiveness on the model’s classification accuracy.

**Accuracy Assessment:** The impact of each weight adjustment is assessed by recalculating the classification accuracy. Each individual  $i_{\text{class}}$  is classified based on the highest similarity score for the networks  $G_1$  and  $G_2$ .

Each individual  $i_{\text{class}}$  in the dataset has a ‘Ground Truth’ label, denoted by  $\tau(i_{\text{class}})$ , which indicates whether the individual is ‘disabled’ or ‘non-disabled’. After all individuals have been classified, we evaluate the accuracy of the model by determining the fraction of individuals that have been correctly classified according to the ‘Ground Truth’. The accuracy of the

Table 1: Features in the HRS Dataset That Have Been Used in This Study.

Variables
Years of education
Ever had cancer
Body Mass Index (BMI)
Ever drinks alcohol
Ever had high blood pressure
Total of all assets
Ever had lung disease
Ever had cancer
Ever had arthritis
Any difficulty-Using the toilet
Any difficulty-Walk across room
Any difficulty-Dressing
Any difficulty-Bathing or showering
Any difficulty-Eating
Any difficulty-Get in/out of bed

Table 2: Experimental Results Summary.

Experiment	TP	TN	FP	FN	FPR	FNR	Accuracy (%)
1	25	31	18	21	0.37	0.46	58.95
2	27	34	16	22	0.32	0.50	61.62
3	25	30	19	23	0.39	0.48	56.70
4	30	35	14	20	0.29	0.40	65.66
5	29	34	16	20	0.32	0.41	63.64
6	27	31	19	21	0.38	0.44	59.18
7	28	38	12	21	0.24	0.43	66.67
8	30	38	12	19	0.24	0.39	68.69
9	25	31	19	25	0.38	0.50	56.00
10	28	37	11	20	0.23	0.42	67.71
Average	27	34	16	21	0.32	0.44	62.48

classification model is calculated as follows:

$$\text{Accuracy} = \frac{1}{N} \sum_{j=1}^N (\text{Classification}(i_{\text{class}_j}) = \tau(i_{\text{class}_j})),$$

where  $N$  is the number of classified nodes,  $i_{\text{class}_j}$  is the node being classified,  $G_1$  and  $G_2$  are the two networks.  $\text{Classification}(i_{\text{class}_j})$ , and  $\tau(i_{\text{class}_j})$  are the predicted classification outcome and the 'Ground Truth' label for the  $j$ -th individual, respectively.

At each iteration, we increased the weight of a single feature, evaluating the impact of this change on the model's accuracy. Then the following steps are applied:

- If the accuracy of the model increases following the weight change, we update the selected weight with the change made. Then we randomly select a weight feature to further explore potential improvements in model performance.
- If there is no improvement in accuracy the weight

is reverted to its previous value, and the adjustment process randomly select a feature different from the one that has been considered in this iteration.

This process is repeated for all features, until the method converges, that is each weight feature change does not improve accuracy.

## 4 EXPERIMENTAL RESULTS

In this section, we present the outcomes of a series of preliminary experiments to evaluate the performance of our predictive model. For each experiment, a distinct random sample of 100 individuals was selected from a larger dataset, which included 50 disabled individuals and 50 non-disabled individuals. It is important to note that some selected individuals (at most 5%) may have incomplete information available in the dataset, hence they are not used for the validation

phase.

We use RAND U.S. Health and Retirement Study (HRS) data <sup>4</sup>. The used dataset comprises health status and risk factor details from 42,406 survey participants born between the years 1890 and 1995. The features in the HRS dataset that were used in this research are described in Table 1.

Table 2 presents the performance of our methods for the data random samples from the test dataset. The performance metrics considered include True Positives (TP), True Negatives (TN), False Positives (FP), False Negatives (FN) <sup>5</sup>, the False Positive Rate (FPR), the False Negative Rate (FNR), and the overall accuracy in percentage. The FPR and FNR provide insights into the model's tendency to categorize negative and positive cases erroneously, which are respectively calculated as:  $FPR = FP / (FP + TN)$ , and  $FNR = FN / (TP + FN)$ . Finally, accuracy quantifies the percentage of actual findings in the dataset that match the ground truth.

In Table 2, experiment 10, which has the highest accuracy at 68.69%, shows a balance between identifying true positives and true negatives while minimizing both false positives and false negatives. In contrast, Experiment 5 shows the lowest accuracy, indicating a higher misclassification rate. On average, these experiments have the accuracy 62.48%, and across the 10 experiments, the model achieved a TP rate of 27, a TN rate of 34, with FP and FN averaging at 16 and 21, respectively. The average FPR was observed at 0.32, with the FNR at 0.44.

In Table 2, a notable pattern across all experiments is the higher number of TN compared to TP, and FN compare to FP. This trend shows that the model has a tendency to classify individuals as 'not-disabled'. In particular, the methods has better performances in correctly identifying individuals who are not disabled than it is at identifying those who are disabled.

## 5 CONCLUSION

This preliminary study explores feature weight optimization for disability classification and shows how learning and network approaches can be integrated into healthcare frameworks in a potentially fruitful way. We plain to compare the results of our method with other prediction methods. Another possible fu-

<sup>4</sup><https://hrs.isr.umich.edu>.

<sup>5</sup>The TP refers to when an individual's ground truth is 'not disabled', but they are incorrectly classified as 'disabled', and the FN refers to when an individual's ground truth is 'disabled', but they are incorrectly classified as 'not disabled'.

ture direction is to improve the ability to classify 'disabled' individuals. Extending our dataset to include a wider variety of demographic and geographic characteristics is expected to enhance the generalizability and relevance of our findings.

## REFERENCES

- Bondy, J. A. and Murty, U. S. R. (2008). *Graph theory*. Springer Publishing Company, Incorporated.
- Cui, H., Lu, J., Wang, S., Xu, R., Ma, W., Yu, S., Yu, Y., Kan, X., Ling, C., Ho, J., et al. (2023). A survey on knowledge graphs for healthcare: Resources, applications, and promises. *arXiv preprint arXiv:2306.04802*.
- Health and Study, R. (2008). Public use dataset. produced and distributed by the university of michigan with funding from the national institute on aging (grant number nia u01ag009740).
- Hosseinzadeh, M. M. (2020). Dense subgraphs in biological networks. In *International conference on current trends in theory and practice of informatics*, pages 711–719. Springer.
- Hosseinzadeh, M. M., Cannataro, M., Guzzi, P. H., and Dondi, R. (2022). Temporal networks in biology and medicine: a survey on models, algorithms, and tools. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 12(1):10.
- Li, Z., Shao, A. W., and Sherris, M. (2017). The impact of systematic trend and uncertainty on mortality and disability in a multistate latent factor model for transition rates. *North American Actuarial Journal*, 21(4):594–610.
- Pham, T., Tao, X., Zhanag, J., Yong, J., Zhang, W., and Cai, Y. (2018). Mining heterogeneous information graph for health status classification. In *2018 5th International Conference on Behavioral, Economic, and Socio-Cultural Computing (BESCC)*, pages 73–78. IEEE.
- Pham, T., Tao, X., Zhang, J., Yong, J., Li, Y., and Xie, H. (2022). Graph-based multi-label disease prediction model learning from medical data and domain knowledge. *Knowledge-based systems*, 235:107662.
- Rossetti, G. and Cazabet, R. (2018). Community discovery in dynamic networks: a survey. *ACM computing surveys (CSUR)*, 51(2):1–37.
- Stuck, A. E., Walthert, J. M., Nikolaus, T., Büla, C. J., Hohmann, C., and Beck, J. C. (1999). Risk factors for functional status decline in community-living elderly people: a systematic literature review. *Social science & medicine*, 48(4):445–469.
- Tao, X., Pham, T., Zhang, J., Yong, J., Goh, W. P., Zhang, W., and Cai, Y. (2020). Mining health knowledge graph for health risk prediction. *World Wide Web*, 23:2341–2362.
- Wang, T., Qiu, R. G., Yu, M., and Zhang, R. (2020). Directed disease networks to facilitate multiple-disease risk assessment modeling. *Decision Support Systems*, 129:113171.