Journal of the American Medical Informatics Association, 26(11), 2019, 1263–1271
doi: 10.1093/jamia/ocz070
Advance Access Publication Date: 14 June 2019
Research and Applications

Research and Applications

Network context matters: graph convolutional network model over social networks improves the detection of unknown HIV infections among young men who have sex with men

Yang Xiang,[1] Kayo Fujimoto,[2] John Schneider,[3,4] Yuxi Jia,[1,5] Degui Zhi,[1] and Cui Tao[1]

[1]School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, Texas, USA, [2]Department of Health Promotion & Behavioral Sciences, School of Public Health, University of Texas Health Science Center at Houston, Houston, Texas, USA, [3]Departments of Medicine and Public Health Sciences, University of Chicago, Chicago, Illinois, USA, [4]Chicago Center for HIV Elimination, University of Chicago, Chicago, Illinois, USA, and [5]Department of Medical Informatics, School of Public Health, Jilin University, Jilin, China

Corresponding Author: Cui Tao, PhD, School of Biomedical Informatics, University of Texas Health Science Center at Houston, Suite 600, 7000 Fannin Street, Houston, TX 77054, USA (Cui.Tao@uth.tmc.edu)

Received 3 October 2018; Revised 11 April 2019; Editorial Decision 25 April 2019; Accepted 28 April 2019

ABSTRACT

Objective: HIV infection risk can be estimated based on not only individual features but also social network information. However, there have been insufficient studies using n machine learning methods that can maximize the utility of such information. Leveraging a state-of-the-art network topology modeling method, graph convolutional networks (GCN), our main objective was to include network information for the task of detecting previously unknown HIV infections.

Materials and Methods: We used multiple social network data (peer referral, social, sex partners, and affiliation with social and health venues) that include 378 young men who had sex with men in Houston, TX, collected between 2014 and 2016. Due to the limited sample size, an ensemble approach was engaged by integrating GCN for modeling information flow and statistical machine learning methods, including random forest and logistic regression, to efficiently model sparse features in individual nodes.

Results: Modeling network information using GCN effectively increased the prediction of HIV status in the social network. The ensemble approach achieved 96.6% on accuracy and 94.6% on F1 measure, which outperformed the baseline methods (GCN, logistic regression, and random forest: 79.0%, 90.5%, 94.4% on accuracy, respectively; and 57.7%, 80.2%, 90.4% on F1). In the networks with missing HIV status, the ensemble also produced promising results.

Conclusion: Network context is a necessary component in modeling infectious disease transmissions such as HIV. GCN, when combined with traditional machine learning approaches, achieved promising performance in detecting previously unknown HIV infections, which may provide a useful tool for combatting the HIV epidemic.

Key words: HIV, epidemiology, machine learning, graph convolutional networks, social networks

INTRODUCTION

According to the Centers for Disease Control and Prevention (CDC), there were about 38 500 new human immunodeficiency virus (HIV) infections in 2015 and over 39 700 in 2016 in the US.[1] From 1981 to 2015, over 1.8 million people in the US have been infected with HIV and approximately 650 000 people have died.[2] The group at the highest risk of HIV diagnosis is men who had sex with men (MSM) (30.7% in 2015). Among the new HIV infections in 2016, 37.2% were in people between 20 and 29 years old, which is the most-affected age group. HIV is costly with about $29.5 billion spent in 2014[3] and $32.8 billion in 2017 in the US.[4] A global public health priority is to develop new methods that identify risk populations most vulnerable to HIV infection and subsequently develop interventions in order to reduce potential HIV transmissions. Among the risk populations, certain proportions of individuals are not aware of their infection status, even though they frequently expose themselves to risk environments.

The aim of this study was to identify individuals within a social network who are at high risk of HIV infection by making predictions of their HIV infection status. This will assist with research into disease prediction and risk evaluation. Using machine learning or other predictive modeling methods, direct analytical approaches have been used to detect HIV infections. Among existing methods, 1 of the most relevant branches is machine learning-based prediction, which is aimed at categorizing the HIV status of an individual based on information on the individuals' features such as demographic characteristics.[5] Another branch is statistical methods, which evaluate risk factors related to the prevalence of HIV such as an individual's socioeconomic status,[6] lack of access to health care,[4] residential environment (eg, urban vs rural), marital status, and sexual behaviors.[7]

However, most of these previous studies lack explorations of network contextual features such as the connections between individuals. In particular, risk environments[8] that comprise sexual networks, social networks,[9,10] and venue-based affiliation networks[11,12] have been shown to play an important role in shaping the HIV/sexually transmitted infection risk. A network study did conduct multi-nomial regression analysis to predict the HIV/syphilis infection status using the degree to which individuals are exposed to network members who are coinfected, HIV mono-infected, or syphilis mono-infected along with individual features.[13] However, that study was limited by a general lack of modeling information flow between individuals.

Stochastic network simulation methodology for modeling HIV transmission dynamics has also been employed. This method uses aggregated empirical egocentrically sampled data as model statistics.[14,15] Such a network modeling approach has the capability of modeling partnership formation and dissolution to better understand the spread of disease in relation to network features such as concurrency and assortative mixing patterns.[15] However, this approach focuses on estimation and statistical inference using the exponential–family random graph models, and their primary focus is not to predict an individuals' disease status.

To address the previously mentioned issues, our study introduces a machine learning method that is based on the social network perspective, which takes into consideration both individual features and network context for the detection of unknown HIV infections. The method was experimented on a data set derived from a multisite longitudinal network study, the Young Men's Affiliation Project (YMAP) in Houston, collected between 2014 and 2016. In that study, only the network built at a single time (the baseline survey) was used. Graph convolutional networks (GCN) is a deep learning method specifically designed for network structures.[16] Rooted in its recent success on graph-based node classifications, and according to the limited graph size, we proposed a novel ensemble approach to predict the HIV status, which integrates GCN into the combination of 2 popular statistical machine learning methods: Random Forest (RF) and Logistic Regression (LR). GCN was mainly used to capture network information such as information flow along edges, while RF and LR showed their advantages in dealing with individual-level sparse features given the limited sample size. Examining these multiple approaches within the same data source can yield important information on which approach is most accurate and the extent to which network data are useful in identifying unknown HIV infections, thus providing a new tool for combatting the HIV epidemic.

MATERIALS AND METHODS

Population

YMAP was a prospective cohort study that examined the impact of social networks formed through social and health venue affiliation on HIV risk and prevention among young men who had sex with men (YMSM), aged 16 to 29, in Houston and Chicago. The present study used the data that were collected in Houston from YMSM participants between 2014 and 2016 through the respondent-driven sampling (RDS) method.[17,18] The RDS method is based on the link-tracing chain referral recruitment method that has been widely employed to recruit hard-to-reach populations such as MSM or drug users. In RDS, individuals were purposively selected as "seeds." These seeds then recruited up to 4 of their contacts (or recruits). In Houston, a total of 378 YMSM were recruited. Survey data was collected based on computer-assisted personal interviews that include sociodemographic characteristics, HIV/sexually transmitted infection risk/protective behaviors, social and sexual networks, and venue attendance or affiliation information.

Determination of HIV and syphilis infection status

Biological data were also collected, and this study used the test results for HIV and syphilis infections based on the Alere Determine HIV-1/2 Combo antigen/antibody test for HIV infection. Participants with reactive samples were confirmed using HIV-1/HIV-2 multi-spot differentiation and HIV RNA (viral load) tests. HIV seropositivity was defined based on the confirmatory test results. Syphilis infection was assessed via a rapid plasma reagin (RPR) test, followed by a confirmatory fluorescent treponemal antibody absorption (FTA) test, and we defined those with FTA test positive as syphilis seropositivity.

Constructing the social network data

To summarize, the aggregated social network data were constructed based on the following 2 sources of network information: (1) survey data in which each participant was asked to nominate up to 5 partners with whom they share personal information (social network) and up to 5 people with whom they had anal, oral, or vaginal sex within the past 6 months (sexual network); (2) peer-referral network that was generated by the RDS recruitment process. These distinct network data sources were combined using a matching procedure based on a fuzzy matching algorithm that cross-refers participants and their partners' sociodemographic information, such as name, age, and race, to determine if the pairs listed are the same persons.[19]

More detailed information on data collection and survey items can be found elsewhere.[13,20]

Incomplete networks simulate real world scenarios

We also built incomplete networks to simulate the scenarios of missing labels. Even if the whole network topology is known, it may contain individuals with unrecorded HIV status. Any proposed method has to be able to handle missing data.

Final network

Finally, we built a social network in which each node stands for an individual participant, and the edges between them represent their social network members with connections. The network contained 378 nodes (positive: 115, negative: 263) and 398 edges.

For each node (individual participant) in the social network, a feature set was built based on the surveys including sociodemographic characteristics, sexual behavioral characteristics, biological values from test results, network characteristics, perceived information, and graph features; and then each feature was converted into numeric values. The graph features were mainly influenced by several popular graph-based features.[21,22] The features adopted within this study are listed in Table 1 (More detailed information is found elsewhere.[23]).

Institutional review board approval

The University of Texas Health Science Center at Houston (UTHealth) received approval from the institutional review board.

Overview of the method

The goal of this study was to determine (with as high a probability as possible) each individual's HIV status (label), when the HIV status of other network members was known. We used statistical machine learning models, such as LR and RF, to extract useful information from individual features and GCN to help model the information flow between individuals. Due to the limited sample size which may hinder the performance of GCN, we used an ensemble approach which combines GCN, LR, and RF with their predicted labels and the corresponding probabilities for each category as its input features.

Graphic convolutional networks

Related studies

While data with regular grid structure (eg, images) can be successfully modeled by convolutional neural networks[24] and sequential structure (eg, natural language) by recurrent neural networks,[25] network data have only recently been explored in the deep learning context.[26] Graph neural networks (GNNs)[26,27] have exploited the idea of information flow along edges, which aim to directly encode the graph structure. They include more comprehensive and thorough network information than previous graph modeling methods, which may destroy the relationships between data samples by using only summary information such as node degree and centrality.[21,22] In the family of graphic neural networks, GCNs (a generalization of convolutional neural networks in the graph Laplacian domain) have been shown to provide promising results on modeling chemical structure of molecules.[28] Kipf and Weilling[16] have shown that a

Table 1 (a). Features for HIV status prediction

Feature	Feature name	Data type and explanation
Sociodemographic characteristics	Age	Continuous
	Race/Ethnicity	Nominal, Hispanic, Non-Hispanic White/Caucasian, Non-Hispanic Black/African American, Non-Hispanic other races
	Education	Nominal, Education level
	Lifetime Homeless	Binary
	Insurance type	Nominal
Sexual behavioral characteristics	Inconsistent condom use	Binary, "not always using condom" with at least 1 sex partner in the past 6 months
	Number of sex partners	Numeric, numbers of sex partners in the past 6 months
	PrEP condom	Ordinal, when taking PrEP, how often did you use condoms during anal or vaginal sex
	Sex transaction	Binary, engaged in sex transaction with at least 1 sex partner in the past 6 months
Biological Data	Viral load	Ordinal, the viral load for each participant
	Syphilis infection	Binary, FTA syphilis test result
Network characteristics	Number of health venues attended	Numeric, total number of health venues* attendance
	Number of social venues attended	Numeric, total number of social venues attendance
	Number of nom sex	Numeric, number of nominated sex partners
	Number of nom soc	Numeric, number of nominated social partners
Information from sampling	Perceived HIV positive	Numeric, number of nominated sex partners perceived as HIV positive by respondents
Graph features	Centrality	Numeric, the eigenvector centrality of a node
	Ratio of positive neighbors	Numeric, number of known HIV positive neighbors (in the training set) in the graph normalized by the number of neighbors
	Ratio of negative neighbors	Numeric, number of known HIV negative neighbors (in the training set) in the graph normalized by the number of neighbors

Abbreviation: FTA, FLUORESCENT TREPONEMAL ANTIBODY-ABSORPTION (FTA-ABS)

*venues: YMSM-serving venues of various types including social (eg, bars, religious or sporting organizations, homeless shelters) and health (eg, clinics, HIV-testing centers).

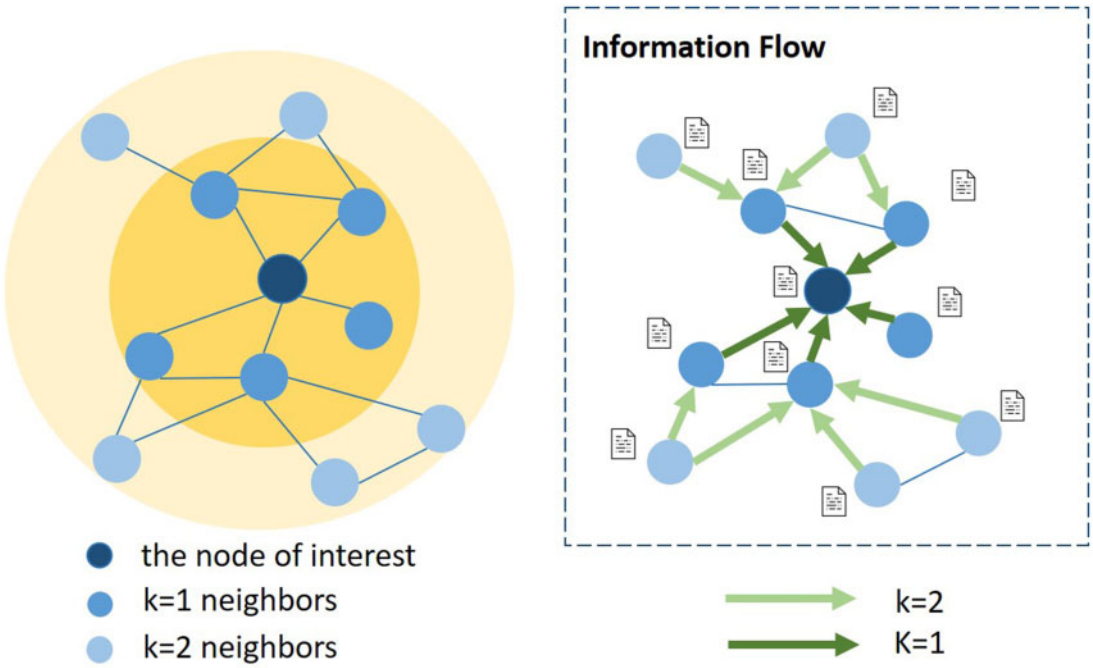

Figure 1. An overview of GCNs in modeling information flow. k = 1 denotes the 1st-order neighbors, and k = 2 denotes the 2nd-order neighbors of the node of interest.

first-order approximation of ChebNet[29] with 2 layers was able to achieve state-of-the-art results in node classification tasks in several networks such as the literature citation network. Since then, GCN approaches have been adopted rapidly to various network-based prediction tasks, including edge prediction,[30] bi-partite graph prediction, and multi-graph prediction.[29] A brief illustration of GCNs in modeling message passing is shown in Figure 1, where the information for a specific node is the aggregation of its neighborhood nodes. After training, a hidden representation for each node is obtained, which can be used for further processing, such as node classification and link prediction.

The GCN model

GCN aims to model each node in a graph using deep neural networks. Let X stand for the N nodes in a graph G, each with a C-dimensional feature set x_i, A is the adjacency matrix of G. A more formal representation of G using a layer-wise propagation rule is outlined as follows:

$$H^{(l+1)} = \sigma(AH^{(l)}W^{(l)}) \tag{1}$$

where $H^{(l)} \in \mathbb{R}^{N\times L}$ is the matrix of activation in the l^{th} layer and $H^{(0)} = X$ stands for the input nodes. $\sigma(\cdot)$ is the activation function such as ReLU.[31] $W^{(l)}$ is the trainable neural network weighted matrix for layer l. The advantage of using Eqation 1 to model each node over conventional graph features is that it aggregates all the neighborhood information through matrix multiplication (the convolution operation) and thus builds a more complete context for each node. The number of convolutional layers (ie, the number of propagations) defines how deep we want GCN to model the neighbors for each node. For example, the first layer stands for modeling the direct neighbors (1^{st}-order) and the second layer stands for modeling the 2^{nd}-order neighborhood, the case of which is shown in Figure 1 using k = 1 (1^{st}-order) and k = 2 (2^{nd}-order).

After adding self-loop to include the information of each node itself and using the diagonal node degree D to normalize the feature vectors, according to Kipf & Welling,[16] the propagation rule can be reduced to

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right) \tag{2}$$

where $\tilde{A} = A + I_N$ is the adjacency matrix of G with added self-connections. I_N is the identity matrix, and $\tilde{D}_{ij} = \sum_j \tilde{A}_{ij}$. It has been validated that this propagation rule can be motivated via a first-order approximation of localized spectral filters on graphs[29] by defining the convolution operation of a signal x with a filter $g_{\theta'}$ as in

$$g_{\theta'} * x \approx \sum_{k=0}^{K} \theta'_k T_k(\tilde{L})x \tag{3}$$

with $\tilde{L} = \frac{2}{\lambda_{max}}L - I_N$ and $L = I_N - \tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}} = U\Lambda U^T$ is the normalized graph Laplacian. In the definition of L, U is the matrix of eigenvectors and Λ is the diagonal matrix of L. When K is limited to 1 and $\lambda_{max} \approx 2$, given the consideration that the neural network can adapt a numeric value during training, the convolution operation can be converted to

$$\begin{aligned} g_{\theta'} * x &\approx \theta'_0 x + \theta'_1(L - I_N)x = \theta'_0 x - \theta'_1 D^{-\frac{1}{2}}AD^{-\frac{1}{2}}x \\ &\approx \theta\left(I_N + D^{-\frac{1}{2}}AD^{-\frac{1}{2}}\right)x \end{aligned} \tag{4}$$

where θs are weight matrices. After constraining the number of parameters θ to avoid overfitting, the convolutional signal matrix for an input X can be further simplified to

$$Z = \hat{A}X\Theta = \tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}X\Theta \tag{5}$$

where $\Theta \in \mathbb{R}^{C\times F}$ is the parameter matrix with a C-dimensional feature vector for each node and F filters for the convolution operation. Rewriting Equation 5 into a function, and if including 2 layers, the mathematical inference of GCN is

$$Z = f(X, A) = \text{softmax}(\hat{A}(\sigma\left(\hat{A}XW^{(0)}\right))W^{(1)}) \quad (6)$$

In Eqation 6, $W^{(0)}$ is the weight matrix connecting the input X and the hidden layer H, and $W^{(1)}$ is the weight matrix connecting H and the outputs. Since $\hat{A}$ can be computed in advance, according to Kipf & Welling,[16] the computational complexity of Eqation 5 is only $O(|\epsilon|CHF)$ (ie, linear in the number of edges in the graph) which is very time efficient. The losses of GCN are cross-entropies based on the node labels, and when adapting itself to graph semi-supervised learning, the losses only include those from nodes with known labels. The model will output a predicted label for each category as well as predicted probabilities.

Statistical machine learning models

LR and RF were employed as the benchmark statistical machine learning classifiers. LR is a popular and effective statistical machine learning model[32] and often employed as a baseline for predictive modeling.[33] It performs well especially when the data set is small and with rich sparse features. RF is an ensemble of decision trees with different subsets of features or training samples and is also a strong baseline for classification.[34] LR and RF take sparse features as inputs and output both predicted probabilities and labels for each node. For LR and RF, the context information was also included in the feature set (we include sparse graph features in Table 1 as a comparison).

Ensemble approach

Statistical machine learning methods are effective in individual feature selection especially with rich sparse features and on small data sets, while GCN is capable of capturing network information such as information flow, but the performance may be hindered by the graph size. Based on the supposition that these 2 types of information may complement each other, we used an ensemble learning approach by integrating GCN into RF and LR. Both the predicted labels and the probabilities for each category were used as the input features of the ensemble classifier, in which the probabilities were obtained from running the classifier on either a training node (with known label) or a testing node (with unknown label). Firstly, each basic classifier was trained separately on the training set, and the predicted probabilities and labels for each node produced by each classifier were extracted respectively. Then, the ensemble model took them as input features and learned to put the data through further operations. In our study, we employed RF as the ensemble classifier so that the best combinations of features were automatically selected and optimized. An overview of the ensemble method is shown in Figure 2. We also compared other ensemble combinations in the experiments.

EXPERIMENTS

Experiment configuration

We set the training: testing ratio as 300: 78 and performed 10 rounds of random resampling. In each round, to conduct semi-supervised learning, N% (with an interval of 10% from 10% to 100%) training nodes with labels were randomly selected 10 times. All the results reported in this section were averaged across the 10 times and 10 rounds random selection. The primary evaluation metrics were accuracy and F1 measure. Accuracy was computed according to the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) in classification. The equations are:

$$TruePositiveRate = TP/(TP + FN) \quad (7)$$

$$FalsePositiveRate = FT/(FP + TN) \quad (8)$$

$$Accuracy = (TN + TP)/(TN + FN + TP + FP) \quad (9)$$

Accuracy was the commonly used measure in previous HIV-related predictions (such as those by Betechuoh et al,[35] Leke-Betechuoh et al,[36] and Dom et al[37]) to evaluate the performance. In addition to accuracy, we also used F1 measure as another primary evaluation metric, which is computed by

$$precision = TP/(TP + FP) \quad (10)$$

$$recall = TP/(TP + FN) \quad (11)$$

$$F1 = 2 \times precision \times recall/(precision + recall) \quad (12)$$

Precision is the percentage of the correctly classified HIV positives, while recall denotes how complete the HIV positives can be identified, and F1 is the balance of the 2.

For the individual classifiers, LR and RF were implemented with the help of the Scikit-learn toolkit for Python3,[38] and the TensorFlow code base from Kipf & Welling[16] was used for the implementation of GCN. In each round of resampling, the testing set was held out as blind. We tuned the parameters of each method over the resampling so that they obtained the optimal results on the training set in the label-complete network (with 100% known training labels). Finally, the parameters for LR and RF were set as: C = 0.01 with L2 penalty for LR, max depth = 3 and random state = 40 for RF. For GCN, the dimension of hidden layers was set to 16 and 2 convolutional layers were used (considering 1st- and 2nd-order neighbors). The model was trained using Adam[39] with the maximum number of iterations at 200 and an early stopping with a window size of 10. The dropout rate for each convolutional layer was set as 0.1 so that more features were retained. For a fair comparison between different classifiers, no thresholding strategies were applied on the predicted probabilities; instead, the default cutoff of 0.5 was used to generate the predicted labels. The ensemble methods evaluated in our experiments include GCN+LR, GCN+RF, LR+RF, and GCN+LR+RF, in which for LR and RF, graph features were removed since we found that removing these features improved the performance of the ensemble methods. To further validate the effectiveness of GCN in modeling network contextual information, a comparison between classifiers with different contextual features (ie, ratios of positive and negative neighbors and eigenvector centrality listed in Table 1 vs GCN) was performed.

RESULTS

The performance of different classifiers for predicting HIV status are shown in Table 2. The results were generated on the label-complete network. We noticed that the ensemble GCN+RF achieved the highest accuracy, and F1 and GCN+LR+RF performed comparably well. Additionally, both GCN+LR and GCN+RF demonstrated considerable improvements over simply using LR and RF, indicating the positive contribution of context factors modeled by GCN. LR+RF also achieved good results compared with individual classifiers. All the ensemble methods had both high accuracies and F1s. Using RF alone can result in a satisfying performance of 90.4% on F1.

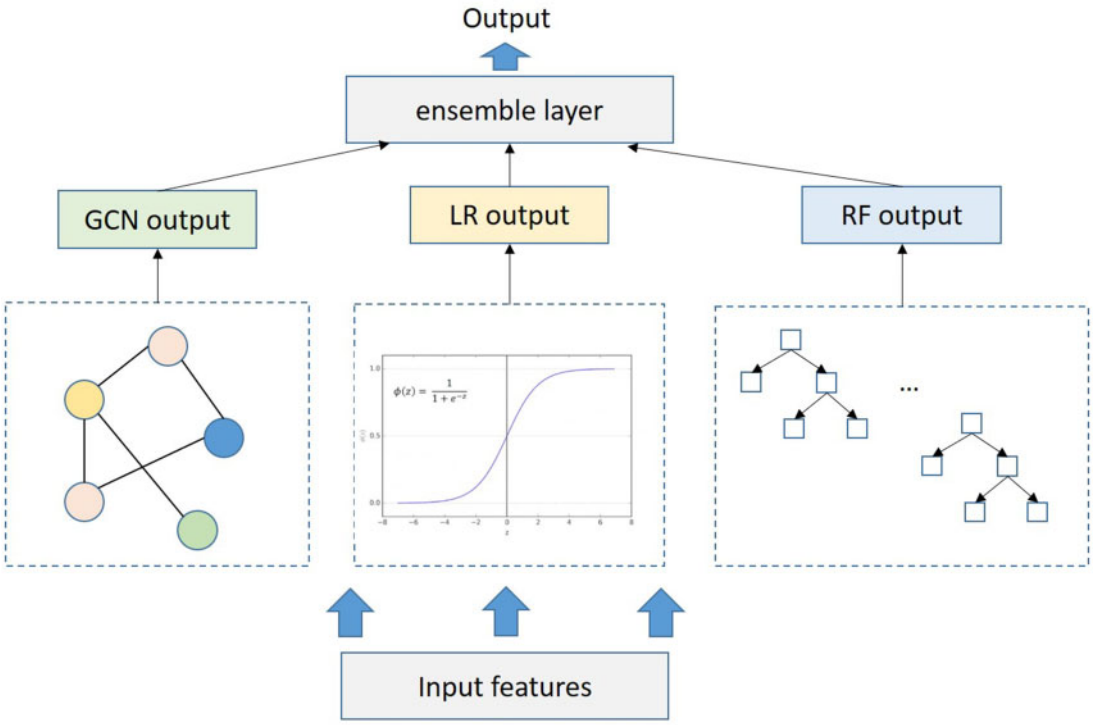

Figure 2. Overview of the ensemble approach.

Table 2. Prediction performance for different classifiers in percent (standard deviation of 10 rounds of resampling)

Method	Accuracy(%)	F1(%)
GCN	79.0(±2.72)	57.7(±6.83)
LR	90.5(±2.69)	80.2(±7.48)
RF	94.4(±2.76)	90.4(±4.38)
GCN+LR	93.4(±3.09)	88.4(±5.30)
GCN+RF	**96.6**(±1.97)	**94.6**(±2.88)
LR+RF	95.3(±2.75)	91.6(±5.64)
GCN+LR+RF	96.5 (±2.05)	94.5(±3.12)

Abbreviations: GCN, graph convolutional networks; LR, logistic regression; RF, random forest. The optimal values for each column are marked bold.

Figures 3 and 4 depict accuracies and F1s for graph semi-supervised learning respectively, where different colors and markers were used to denote different methods. Accuracy and F1 curves basically follow the same trend: the curves generated by the ensemble methods achieved higher accuracy than only using the basic methods, and the performance improves with increased fraction of labeled training nodes. The ensemble GCN+RF and GCN+RF+LR produced the optimal results for almost every fraction of labeled training nodes. Among the basic classifiers, RF performed consistently better than LR and GCN. The ensemble LR+RF and GCN+RF also produced comparable performance to GCN+RF+LR. Although GCN itself did not produce as good results as other methods, when adding it to LR and RF, notable improvements were generated.

The ensemble methods (except GCN+LR) had comparable performances and can get over 85% on accuracy and 80% on F1 even with only 20% training labels, indicating that the individual features and context information can compensate well for prediction in our ensemble methods. A Wilcoxon test was performed to test the significance of the improvements. GCN+RF showed an averaged P value of .159 over RF. This only shows a near-marginal significance, signifying that the uncertainty in resampling on such a small network will affect the performance, and that RF alone behaved as a strong baseline. In addition, the difference between GCN+LR and LR is significant as to alpha = 0.05 (0.0192).

The comparison between classifiers with different types of context features is shown in Table 3. The results demonstrate that adding graph features to LR and RF did not improve the performance—and even made the result worse. A possible cause for this decreased performance might be that the summary-like graph features (eg, ratio of positive neighbors) added uncertainties into the feature set. Results for adding graph features under the semi-supervised setting were not satisfying either. For example, LR and RF produced 60.8% and 66.7% on F1 with graph features compared with GCN+LR and GCN+RF which produced 81.2% and 89.6% given 50% labeled nodes.

DISCUSSION

In this study of identifying social network members who are at high risk of HIV seropositivity, but whose status is unknown, we found that the machine learning methods produced satisfying results with the addition of context features modeled by GCN.

GCN in modeling context

LR and RF are both strong baselines for classifications on relatively small data sets. As shown in Table 2 and Figures 3 and 4, GCN generally performed worse than both of them on accuracy and F1. A possible reason might be that as a deep learning model, GCN requires more parameters to train than statistical machine learning

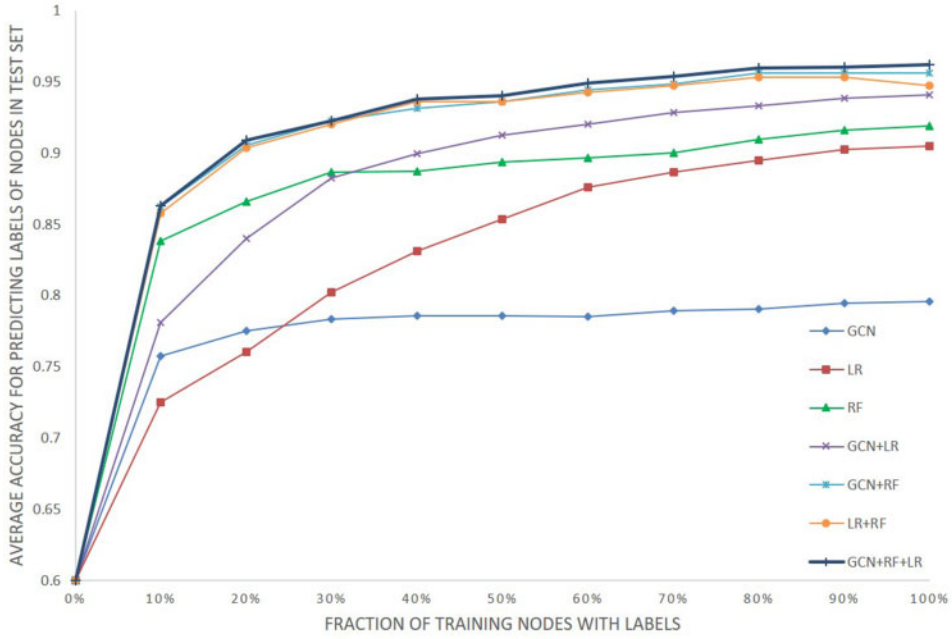

Figure 3. Accuracies produced by different methods in graph-semi supervised learning.

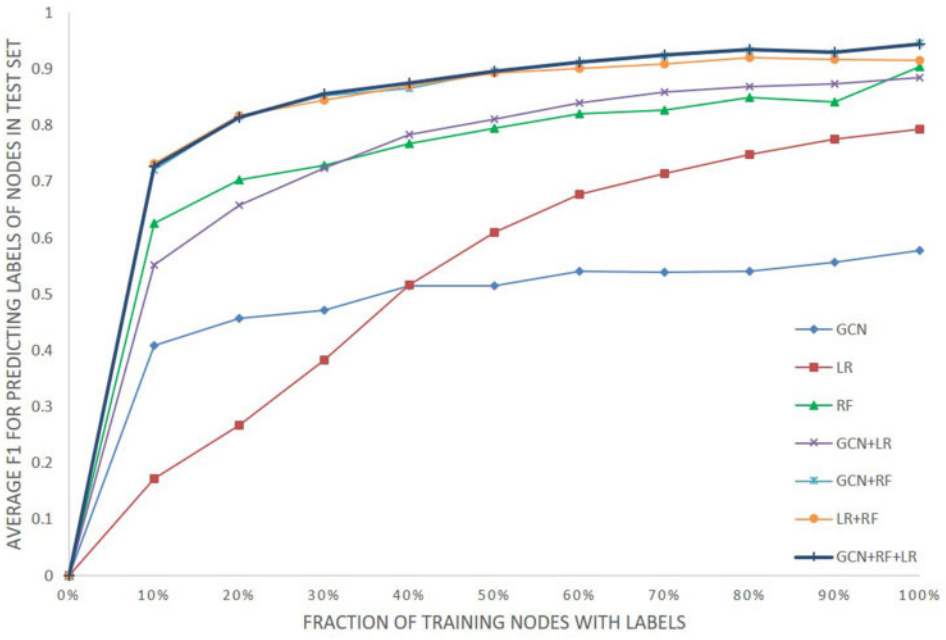

Figure 4. F1s produced by different methods in graph-semi supervised learning.

Table 3. F1s generated by LR and RF considering different types of context information

Method	original	+graph features	+GCN	+graph features +GCN
LR	80.2%	79.3%	88.4%	**91.2**%
RF	90.4%	76.8%	**94.6**%	90.9%

Abbreviations: GCN, graph convolutional networks; LR, logistic regression; RF, random forest. The optimal values for each column are marked bold.

methods. It naturally relies on bigger data sets which cannot be obtained in this network due to the limited size of the cohort. Another potential reason for the lower performance is GCN's insufficiency on feature selection. GCN lacks sparse feature selection strategies such as ensemble of feature subsets that are adopted in. However, in most cases, an effective utilization of the rich features is quite useful for machine learning in smaller data sets. In this study, LR and RF were able to remedy these drawbacks.

Improvements are obvious when GCN is added to LR and RF, demonstrating the important contribution of contextual information modeled by GCN in prediction. LR and RF cannot sufficiently model the network information, unlike GCN, despite having several features related to the context (eg, number of sex partners). Some methods enable label propagations from the observed node to its neighbors; thus, if the observed node is positive, neighbors are likely to be positive as well (eg, LR with neighborhood labels as features). However, since the node label is determined by not only the neighborhood information but also by individual features (eg, condom usage) only if a method is able to model these 2 types of features simultaneously and effectively can it generate a better result, which can be seen from Table 3. In our present study, we did not include any known node labels as features for GCN. This might be part of our future work.

The ensemble

The 4 ensemble methods we tested all demonstrated their superiority in integrating weak classifiers. Nevertheless, the relative improvements of the ensembles (eg, GCN+LR improved LR by over 8%) derived from Table 2 show that GCN plays a more important role in offering complementary effects (ie, adding contextual information). LR+RF, although this combination also performed well, did not boost RF much. The comparisons of RF vs RF+LR, and GCN+RF vs GCN+RF+LR reveal that RF and LR may have a large proportion of consistent predictions, whereas RF performed better than LR since adding LR did not produce obvious improvements in these classifiers. The results in Figures 3 and 4 show that the ensemble approach can help to obtain satisfying predictions even in a network with missing labels.

Limitations and future work

The work still has some limitations at present: 1) The GCN model itself did not perform well, especially on F1, partly due to the small training sample size and its insufficiency on modeling sparse features. 2) Due to the limited graph size, the deviations on the results were big and the improvements over baselines sometimes showed only marginal significance when doing average on resampling (see Table 2), and some methods are very sensitive to parameters (eg, RF). In future studies, we plan to collect additional data and design ways of including other types of edges such as the venue coaffiliation ties between pairs of nodes. Also, we may try to further investigate how to improve the feature extraction part of GCN. Furthermore, this approach can also enhance existing agent-based simulation methods of HIV and other infectious diseases by adding prediction in intermediate steps so as to speed up the computation and make predictions on bigger data sets.

CONCLUSION

This paper presents a novel approach to combine individual features and network contextual information in detecting unknown HIV infections with the framework of integrating social network perspective and machine learning methodologies. And we present innovative approaches combining statistical machine learning methods and graph-based deep learning methods on HIV status prediction. The capacities of distinct classifiers were evaluated under a semi-supervised learning paradigm, which was a simulation of missing recorded HIV status in real-world HIV networks. Experimental results validated the effectiveness of the ensemble approach in networks with different fractions of known labels. The ensemble methods produced promising results on HIV infection detection and are expected to provide useful clinical support for HIV prevention.

FUNDING

This work was supported by the National Institutes of Health, grant nos. 1R01MH100021, 1R01DA039934, 1R01AI130460, and R01LM011829.

AUTHOR CONTRIBUTIONS

DZ and CT conceived and supervised the study. YX and DZ lead the design of the model and experiments. YX conducted the experiment and wrote the manuscript. KF conducted the literature review, collected the data, and helped to do feature selections. JS helped to modify the literature review, polished the manuscript, and provided professional background knowledge on interpretations. YJ was in charge of the display of all the figures and representations. DZ and CT polished the manuscript. All authors approved the final draft.

ACKNOWLEDGMENTS

We thank Dr Irmgard Willcockson for editorial help. We also acknowledge the support of Nvidia Corporation and Texas Advanced Computing Center for providing the computational resources.

CONFLICT OF INTEREST STATEMENT

None declared.

REFERENCES

1. Centers for Disease Control and Prevention (CDC). HIV in the United States: at a glance. http://www.cdc.gov/hiv/statistics/overview/ataglance.html. Accessed October 3, 2018.
2. CDC. Basic statistics. https://www.cdc.gov/hiv/basics/statistics.html. Accessed October 3, 2018.
3. Jones DM. The impact of HIV and AIDS in the United States. http://knowledgecenter.csg.org/kc/content/impact-hiv-and-aids-united-states. Accessed October 1, 2018.
4. CDC. Today's HIV/AIDS epidemic. https://www.cdc.gov/nchhstp/newsroom/docs/factsheets/hiv-todaysepidemic-508.pdf. Accessed October 3, 2018.
5. Leke-Betechuoh B, Marwala T, Tim T, *et al*. Prediction of HIV status 90 from demographic data using neural networks. In: 2006 IEEE International Conference on Systems, Man and Cybernetics 2006: 2339–44. IEEE. Taipei, Taiwan.
6. Abrutyn S, Mueller AS. Prediction of HIV sexual risk behaviors among disadvantaged African American adults using a syndemic conceptual framework. *AIDS Behav* 2015; 79: 211–27. doi: 10.1177/0003122413519445.
7. Kim AA, Parekh BS, Umuro M, *et al*. Identifying risk factors for recent HIV infection in Kenya using a recent infection testing algorithm: results from a nationally representative population-based survey. *PLoS One* 2016; 11: 1–21.
8. Friedman SR, Neaigus A, Jose B, *et al*. Sociometric risk networks and risk for HIV infection. *Am J Public Health* 1997; 87 (8): 1289–96.
9. Schneider JA, Cornwell B, Ostrow D, *et al*. Network mixing and network influences most linked to HIV infection and risk behavior in the HIV epidemic among black men who have sex with men. *Am J Public Health* 2013; 103: 28–36.
10. Friedman SR, Aral S. Special feature: social networks social networks, risk-potential networks, health, and disease. *J Urban Health* 2001; 78: 411–8.

11. Fujimoto K, Wang P, Ross MW, *et al*. Venue-mediated weak ties in multiplex HIV transmission risk networks among drug-using male sex workers and associates. *Am J Public Health* 2015; 105 (6): 1128–35.
12. Street WW, Diego S. Using sexual affiliation networks to describe the sexual structure of a population. *Sex Transm Infect* 2007; 83: 37–42. doi: 10.1136/sti.2006.023580.
13. Fujimoto K, Flash CA, Kuhns LM, *et al*. Social networks as drivers of syphilis and HIV infection among young men who have sex with men. *Sex Transm Infect* 2018;94(5):365–71. doi: 10.1136/sextrans-2017-053288.
14. Goodreau SM, Cassels S, Kasprzyk D, *et al*. Concurrent partnerships, acute infection, and HIV epidemic dynamics among young adults in Zimbabwe. *AIDS Behav* 2012; 16 (2): 312–22.
15. Krivitsky PN, Morris M. Inference for social network models from egocentrically sampled data, with application to understanding persistant racial disparities in HIV prevalence in the US. *Ann Appl Stat* 2017; 11 (1): 427–55.
16. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. *ICLR* 2017; 1–11. doi: 10.1051/0004-6361/201527329.
17. Heckathorn DD. Respondent-driven sampling: a new approach to the study of hidden populations. *Soc Probl* 1997; 44 (2): 174–99.
18. Heckathorn DD. Respondent-driven sampling II: deriving valid population estimates from chain-referral samples of hidden populations. *Soc Probl* 2002; 49 (1): 11–34.
19. Laumann EO, Schneider JA. Structural bridging network position is associated with HIV status in a younger Black men who have sex with men epidemic. *AIDS Behav* 2015; 18: 335–45.
20. Fujimoto K, Cao M, Kuhns LM, *et al*. Statistical adjustment of network degree in respondent-driven sampling estimators: venue attendance as a proxy for network size among young MSM. *Soc Networks* 2018; 54: 118–31.
21. Cheng J, Romero DM, Meeder B, *et al*. Predicting reciprocity in social networks. In: Proceedings of the 2011 IEEE International Conference on Privacy, Security, Risk and Trust and IEEE International Conference on Social Computing PASSAT/SocialCom, 2011: 49–56. doi: 10.1109/PASSAT/SocialCom.2011.110.
22. Bonacich P. Some unique properties of eigenvector centrality. *Soc Networks* 2007; 29 (4): 555–64.
23. Fujimoto K, Wang P, Kuhns L, *et al*. Multiplex competition, collaboration, and funding networks among social and health organizations: towards organization-based HIV interventions for young men who have sex with men. *Med Care* 2017; 55 (2): 102–10.
24. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Adv Neural Inf Process Syst* 2012; 1–9. doi: http://dx.doi.org/10.1016/j.protcy.2014.09.007.
25. Sak H, Senior A, Beaufays F. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. *Interspeech* 2014; 338–42. doi: arXiv: 1402.1128.
26. Scarselli F, Gori M, Tsoi AC, *et al*. The graph neural network model. *IEEE Trans Neural Netw* 2009; 20: 61–80.
27. Gori M, Monfardini G, Scarselli F. A new model for learning in graph domains. *Proc Int Jt Conf Neural Networks* 2005; 2: 729–34.
28. Duvenaud D, Maclaurin D, Aguilera-Iparraguirre J, *et al*. Convolutional networks on graphs for learning molecular fingerprints. *Adv Neural Inf Process Syst* 2015; 1–9. doi: 10.1021/acs.jcim.5b00572.
29. Defferrard M, Bresson X, Vandergheynst P. Convolutional neural networks on graphs with fast localized spectral filtering. *Adv Neural Inf Process Syst* 2016; 3844–52.
30. Zhang M, Chen Y. Link prediction based on graph neural networks. Adv Neural Inf Process Syst 2018; 5165–75. doi:https://doi.org/10.3390/info10050172.
31. Nair V, Hinton GE. Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th International Conference on Machine Learning 2010: 807–14. doi: 10.1.1.165.6419.
32. Hosmer DW Jr, Lemeshow S, Sturdivant RX. Applied Logistic Regression. New York: John Wiley & Sons 2013.
33. Choi E, Bahadori MT, Kulas JA, *et al*. RETAIN: an interpretable predictive model for healthcare using reverse time attention mechanism. *Adv Neural Inf Process Syst* 2016; 3504–42.
34. Liaw A, Wiener M. Classification and regression by randomForest. *R News* 2002; 2: 18–22.
35. Leke-Betechuoh B, Marwala T, Tim T, *et al*. Prediction of HIV status from demographic data using neural networks. *Syst Man Cybern* 2006; 3: 2339–44.
36. Betechuoh BL, Marwala T, Tettey T. Autoencoder networks for HIV classification. *Curr Sci* 2006; 91: 1467–73.
37. Dom RM, Kareem SA, Abidin B, *et al*. The prediction of AIDS survival: a data mining approach. In: Proceedings of the 2nd WSEAS International Conference on Multivariate Analysis and its Application in Science and Engineering, vol. 04, 2016: 48–53.
38. Pedregosa F, Varoquaux G, Gramfort A, *et al*. Scikit-learn: machine learning in python. *J Mach Learn Res* 2012; 12: 2825–30.
39. Kingma DP, Ba J. Adam: a method for stochastic optimization. In: *ICLR*, 2015. doi: http://doi.acm.org.ezproxy.lib.ucf.edu/10.1145/1830483.1830503.