IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 20 November 2023, accepted 12 January 2024, date of publication 17 January 2024, date of current version 23 January 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3355096

RESEARCH ARTICLE

Node Significance Analysis in Complex Networks Using Machine Learning and Centrality Measures

KODURU HAJARATHAIAH[1], MURALI KRISHNA ENDURI[2], (Member, IEEE), SATISH ANAMALAMUDI[2], ASHU ABDUL[3], (Member, IEEE), AND JENHUI CHEN[4,5,6], (Senior Member, IEEE)

[1]School of Computer Science and Engineering, VIT-AP University, Amaravati, Andhra Pradesh 522237, India
[2]Algorithms and Complexity Theory Laboratory, Department of Computer Science and Engineering, SRM University-AP, Amaravati 522240, India
[3]Centre for Computational and Integrative Sciences, Department of Computer Science and Engineering, SRM University-AP, Amaravati 522240, India
[4]Department of Computer Science and Information Engineering, Chang Gung University, Guishan, Taoyuan 33302, Taiwan
[5]Division of Breast Surgery and General Surgery, Department of Surgery, Chang Gung Memorial Hospital, Guishan, Taoyuan 33375, Taiwan
[6]Department of Electronic Engineering, Ming Chi University of Technology, Taishan, New Taipei City 24301, Taiwan

Corresponding author: Jenhui Chen (jhchen@mail.cgu.edu.tw)

This work was supported in part by the National Science and Technology Council, Taiwan, under Grant 110-2221-E-182-041-MY3; and in part by the Chang Gung Memorial Hospital, Taoyuan, Taiwan, under Grant CMRPD2N0051.

ABSTRACT The study addresses the limitations of traditional centrality measures in complex networks, especially in disease-spreading situations, due to their inability to fully grasp the intricate connection between a node's functional importance and structural attributes. To tackle this issue, the research introduces an innovative framework that employs machine learning techniques to evaluate the significance of nodes in transmission scenarios. This framework incorporates various centrality measures like degree, clustering coefficient, Katz, local relative change in average clustering coefficient, average Katz, and average degree (LRACC, LRAK, and LRAD) to create a feature vector for each node. These methods capture diverse topological structures of nodes and incorporate the infection rate, a critical factor in understanding propagation scenarios. To establish accurate labels for node significance, propagation tests are simulated using epidemic models (SIR and Independent Cascade models). Machine learning methods are employed to capture the complex relationship between a node's true spreadability and infection rate. The performance of the machine learning model is compared to traditional centrality methods in two scenarios. In the first scenario, training and testing data are sourced from the same network, highlighting the superior accuracy of the machine learning approach. In the second scenario, training data from one network and testing data from another are used, where LRACC, LRAK, and LRAD outperform the machine learning methods.

INDEX TERMS Complex networks, influential nodes, local centralities, machine learning techniques.

I. INTRODUCTION

The rapid growth of the internet and social networks has led to emerging challenges in network analysis, including detecting highly influential nodes and examining information propagation dynamics within complex networks [1]. Addressing the crucial task of identifying pivotal nodes presents numerous practical applications in areas such as advertising, education, and news dissemination [2]. Conventional methodologies for tackling this challenge primarily rely on structural characteristics, including node degree [3], shortest path length between all node pairs [4], average shortest path length between all pairs [5], and node eigenvalues [6], to assess and rank nodes according to their influence. To pinpoint the relationships between nodes and gauge their influence within the network, network structure is crucial. Various approaches based on network structure can be categorized into two groups. The first category focuses on the neighborhood of nodes, including Degree centrality (D) [7], K-shell decomposition [8], and H-index [9], among others.

The associate editor coordinating the review of this manuscript and approving it for publication was Byung Cheol Song.

The second category examines the paths between nodes, such as Eccentricity [10], and Katz centrality (K) [11]. Some of the iterative-based centrality methods are eigenvector [12] and PageRank [13].

In practical scenarios, networks often generate a substantial volume of data, leading to high time complexity and difficulties in measuring centrality. Centrality measures are crucial in identifying critical and influential elements within large datasets [14]. The finding of key nodes is a common challenge in various network applications, including virus spreading [15], image recognition [16], disease propagation [17], [18], spam detection [19], information dissemination [20], and online/offline network activities. Various centralities are proposed to find the vital nodes in complex networks [21], [22], [23].

In a recent study by Lv et al. [24], a novel centrality measure known as relative change of average shortest path (RASP) centrality was introduced, focusing on the overall structure of the network. The RASP method evaluates the change in the average shortest paths across the entire network upon removing a specific node, providing a quantitative measure of information propagation between all pairs of nodes when that node is absent from the network. However, the authors are considered for the entire network, and computational time is also more if the entire network is considered. Later, there has been a proposal for a new centrality measure (Local RASP) that explores information propagation in networks, even considering the removal of nodes with the importance of local structure [25].

Recently, Zhao et al. [26], [27] introduced a data-driven machine-learning approach to identify influential nodes in complex networks. Their method involved employing a classification model for vital node identification and training it on a significant portion of nodes from the original network. Rezaei et al. [28] introduced an innovative sampling technique named cluster sampling, which guarantees the inclusion of nodes with diverse structural and influential properties in the training set. The sampling size is confined to a mere 0.5% of the complete network, leading to compact training sets, even in the case of expansive networks. To tackle the constraint of inadequate training data, they selected a Support Vector Regression machine featuring a Radial Basis kernel as their machine learning model. Nonetheless, generalizing the relative change in local centralities is not a typical strength of machine learning algorithms.

From this motivation, our paper presents a novel approach to assessing the significance of nodes, which differs from conventional methods that typically rely on evaluating node importance based on specific local topologies. Our proposed framework leverages machine learning techniques, transforming the evaluation of node relevance in a classification challenge. Within this framework, each node's feature vector comprises the values of six well-known and established centralities alongside the spread of infection rate in a propagation environment. The assignment of labels to each node is determined by evaluating the actual propagating capability achieved through simulated propagation. We employed seven ML algorithms to extract the rules governing node importance evaluation.

In the testing and training of a network, we apply ML methods by using centrality and infection rate features to find the ML-based and centrality-based accuracy. More specifically, the main contributions of our work are given as follows.

1) We introduce a universal centrality metric by utilizing the relative alteration in local centrality. In this context, we examine the impact of the centrality metric when a vertex is removed.
2) We introduce three novel centrality metrics, namely LRAD, LRACC, and LRAK, and perform a comparative analysis with conventional measures such as D, CC, K, NBA, DT, RF, SVM, KNN, LR, MLP, and SVM combined with k-means. These newly proposed metrics prove to be highly valuable in examining complex networks, offering reduced computational complexity.
3) To validate the extent of information dissemination, we evaluate our centrality metrics on real-world datasets employing SIR and IC models.

The rest of this paper is organized as follows. We give the related work to the study in Section II and the preliminaries in Section III. Section IV describes generalizing the relative change in local centralities. Section V presents the proposed framework, providing detailed information about its components. We describe the experimental data sets used in Section VI. The results section is described in Section VII. The discussion section is described in Section VII-C. Finally, we address the conclusion and extended works in Section VIII.

ABBREVIATIONS

Here is a list of the abbreviations employed in this document.

1) D: Degree centrality
2) CC: Clustering Coefficient
3) K: Katz centrality
4) LRAD: Local Relative change in Average Degree
5) LRACC: Local Relative change in Average Clustering Coefficient
6) LRAK: Local Relative change in Average Katz
7) SIR: Susceptible Infected Recovered
8) NBA: Naive Bayesian
9) DT: Decision Tree
10) RF: Random Forest
11) SVM: Support Vector Machine
12) KNN: K-Nearest Neighbor
13) LR: Logistic Regression
14) MLP: Multi-Layer Perceptron
15) SVM+K-means: Support Vector Machine + k-means

II. RELATED WORK

In this section, we delve into the existing research concerning ranking nodes' influence within complex networks. Over the years, researchers have introduced several classic centrality

techniques aimed at gauging the topological significance of nodes. These methods can broadly be classified into four categories: (a) Neighborhood-based Centrality: This category includes metrics like degree and H-index [3]. These metrics assess a node's importance based on the number of immediate or multi-step neighbors, emphasizing local connections. (b) Path-based Centrality: Metrics like closeness [29] and Katz centrality [11] fall under this category. They measure a node's importance by considering the shortest paths to other nodes, focusing on efficient communication within the network. (c) Iterative Refinement Centrality: Techniques like Eigenvector centrality [6] and LeaderRank [30] belong to this group. The assessment of a node's significance involves appraising the importance of its neighbors through a systematic iteration of the network structure. (d) Node-operation-based Centrality: This category encompasses methods such as connectivity-sensitive approaches [31] and stability-based techniques [32]. These methods evaluate a node's importance by observing the impact on the network's structure when the node is deleted or merged. While these methods effectively identify influential nodes within specific structural confines, they often focus on particular network attributes. Additionally, some studies have employed machine learning techniques to identify influential nodes, especially in context-specific scenarios, such as determining user influence on social platforms like Facebook [33]. Berahmand et al. [34] introduced a novel centrality measure rooted in the inherent features of complex networks. Their innovative approach assigns elevated rankings to structural holes, recognizing them as superior spreaders within the network. This centrality measure utilizes the positive impact of second-level neighbors' clustering coefficients and simultaneously accounts for the negative influence of a node's clustering coefficient in evaluating node importance. As a result, this approach guarantees the identification of spreaders that are not excessively close to each other. The relevant literature on machine learning and centralities is outlined in Table 1.

In this study, we focus on devising centrality measures to pinpoint the vital node that ensures maximum information spread while maintaining minimal time complexity. To achieve this objective, we introduce three novel centrality measures: LRACC, LRAD, and LRAK. These approaches focus on local average structural information. They are derived from generalized centrality measures that assess the relative change in degree, Katz, and clustering coefficient after the removal of a node. Specifically, the local centrality measures are labeled LRACC, LRAD, and LRAK. To evaluate their effectiveness, we compare these proposed methods against standard measures in existing literature, including D, CC, K, and various machine learning algorithms.

III. PRELIMINARIES

One can ascertain the significant importance of nodes in a network by employing centrality metrics. To determine influential nodes, the following are measures that are frequently used: The centrality degree of a node depends on the number of connections it possesses. The symbols and annotations utilized in this work have been outlined in Table 2.

High-degree nodes are seen as having enormous influence. Node i's degree centrality [3] in the network can be calculated using the following formula

$$C_d(i) = \frac{d_i}{(N-1)}, \tag{1}$$

where d_i represents the node's degree i, and N denotes nodes present in the network. The clustering coefficient gauges how interconnected a node's neighbors are. It measures the likelihood that a node's neighbors are also connected to one another, forming a cluster. It is a measure that provides information about the network's local structure. In complex networks, influential nodes often have a high clustering [39], as they tend to connect to other highly connected nodes, forming dense clusters.

A node with a significantly high clustering coefficient indicates it is well connected to its neighbors and is likely to operate as a network influencer node. For a specific given node i, the clustering coefficient is defined as triangles (i.e., fully connected triads) involving node i divided by triangles possibility that could involve node i. Mathematically, it can be represented as

$$C_i = \frac{2T_i}{d_i(d_i-1)}, \tag{2}$$

where T_i denotes the number of triangles centered on the node i and d_i denotes the degree of node i. Every node's relative influence is determined by Katz centrality (K) by considering the immediate neighboring nodes along with the subsequent nodes that are neighboring nodes and connected to the immediate neighbors. The K for a node v_i is known as

$$K(v_i) = \alpha \sum_{j=1}^{n} A_{i,j} K(v_j), \tag{3}$$

where α is a dumping factor, which is considered less significant than the biggest eigenvalue.

Our goal was to identify important nodes in complex networks using diverse methods in machine learning (ML), including Support Vector Machines (SVM) [40], [41], Decision Trees (DT) [42], Logistic Regression (LR) [43], k-Nearest Neighbors (KNN) [44], Random Forests (RF) [45], Naive Bayes (NBA) [46], Multilayer Perceptron (MLP) [47], and a combination of clustering and classification (K-means + SVM). By leveraging these ML algorithms, we can accurately predict and classify the significant nodes within complex networks, providing a streamlined approach to network analysis.

A. SUPERVISED LEARNING

One of the machine learning approaches is Supervised learning [48], where a function is trained to convert inputs to outputs using labeled training data. The function is inferred

TABLE 1. Literature related to machine learning methods and Centralities.

Reference	Contribution	Advantages	Limitations
Zahao et al. [26]	Introduced a machine learning-based framework for assessing the significance of nodes in the context of information propagation.	Based on the labels, spreadability is easy.	No assessment of the influence of node removal on the relative change in ML algorithms.
Hajarathaiah et al. [25]	Proposed a new centrality measure (Local RASP) that explores information propagation in networks, even considering the removal of nodes with the importance of local structure.	Computation is less	No generalized the local centrality.
Hajarathaiah et al. [35]	Introduced metrics that rely on evaluating the proportional shift in clustering coefficient, degree, and Katz centralities following the elimination of a vertex.	Generalized the local centralities	ML algorithms not applied
Rezaei et al. [28]	Proposed a unique sampling method named cluster sampling, designed to incorporate nodes with varied structural and influential characteristics into the training set. The sampling strategy involves selecting only 0.5% of the complete network, resulting in modest training sets, even for extensive networks. To address the constraint of limited training data, a Support Vector Regression (SVR) machine with a Radial Basis (RBF) kernel was adopted as the machine learning model.	The SVR machine with an RBF kernel as their ML model to overcome the limited training data.	No generalized ML algorithms applied.
Alshahrani et al. [36]	The algorithms introduced combine degree centrality as a local metric and Katz centrality as a global measure within the frameworks of Independent Cascade (IC) and Linear Threshold (LT) models, focusing on spreading ability and time complexity.	No generalized ML algorithms applied.	
Grando et al. [37]	Utilizing artificial neural networks for proposing, explaining, testing, and comparing centrality measures, addressing the challenges in computing centrality in large, real-world networks.	Showcasing a generative model's ability to provide unlimited training data and reduce computation costs by 30% in real-world scenarios.	Accuracy and computational resources are limited
Jeyasudha and Usha [38]	This paper introduces a method to (1) detect communities using the Laplacian Transition Matrix and community algorithms based on popular hashtags, (2) identify influential nodes within the community using intelligent centrality measures, and (3) implement machine learning algorithms for classifying user intensity, with extensive experiments conducted on COVID-19 datasets.	The proposed approach demonstrates high accuracy (98.6%) using SVM and PCA compared to linear regression when employing new centrality measures.	Require further refinement for precise differentiation.

TABLE 2. List of notations used in this research work.

Symbol	Description
G	A graph
V	Set of vertices or nodes
E	Set of Edges
N	Number of nodes
d_i	Degree of node i
T_i	Number of triangles centered on the node i
L	Neighborhood level
$N_L(v)$	Set of neighboring vertices of vertex v in the network up to level L
$G_{N_L(v)}$	A subgraph on set of nodes $N_L(v)$
$G_{N_L(v)} \setminus v$	A subgraph after deleting a vertex v from $G_{N_L(v)}$
$LRAvg\mathscr{C}$	Local relative change of average centrality $\mathscr{C}$
$AD[G]$	Average degree of the graph G
$ACC[G]$	Average clustering coefficient of the graph G
$AK[G]$	Average Katz centrality of the graph G

using a set of training instances with corresponding inputs and outputs. The algorithms used in this sort of learning use a dataset divided into training and testing sets and depend on outside assistance. By identifying patterns from the input-output pairs in the training set, the objective is to predict or categorize the output variable in the test set.

1) LOGISTIC REGRESSION (LR)

A predictive analytic technique known as logistic regression [43] addresses issues with machine learning categorization. By calculating the probabilities that fall between 0 and 1, independent factors are used to predict the categorical dependent variable. The outcome of an LR analysis must be discrete or categorical. The method uses a logistic function in the shape of "S" to predict the two highest values instead of a regression line and provides probabilistic values rather than precise values. The classification of various data types, the provision of probabilities, and the speedy identification of the most pertinent factors for classification make logistic regression a significant tool.

2) NAIVE BAYES (NBA)

Webb et al. [46] stated that Naive Bayes is a classification technique that relies on the Bayes theorem and assumes predictor independence. The fundamental tenet of this strategy is that no other features in the class are affected by the presence of one feature. Text categorization typically uses Naive Bayes based on the conditional likelihood of occurrence. It is mostly used to group and categorize things, where

$$P(c|x) = \frac{P(c|x)P(c)}{P(x)}. \tag{4}$$

3) DECISION TREES (DT)

Instances are categorized using Decision Trees [42], which sort instances according to the values of their features. Every individual branch within the tree represents a specific value assigned to the corresponding node, which indicates a feature

in the instance being classified. Instances are organized and categorized by feature values starting from the root node. In decision tree learning, a machine learning technique, observations of an item are linked to predictions of its intended value using a decision tree model. These models can be referred to as regression trees or classification trees. Post-pruning procedures are used to evaluate the effectiveness of the decision trees after they have been pruned using a validation set. Even if a node is removed, the instances can still be sorted and assigned to the most common class.

4) RANDOM FOREST (RF)

By employing bagging to create many decision trees, the Random Forest approach [45] classifies fresh incoming data instances into classes or groups. Instead of being pruned, trees are built. When decision trees are being built, the Gini-index cost function is used to locate the best-split point, hence the name randomly, which is a pick of random n features or attributes. The trees are less correlated and have a reduced error rate because the predictor variables are chosen at random. The newly observed data is shared with all classification trees within the random forest in order to compute the target value for the new data instance. The number of predictions made by each classification tree for each class is counted. When a new data instance is created, it returns and considers the class label with the highest votes.

5) K-NEAREST NEIGHBORS (KNN)

Compared to other algorithms, such as Naive Bayes, Random Forest, and Decision Trees, the k-Nearest Neighbours (KNN) algorithm [44] has a faster training time. One notable difference is that KNN, a lazy-learning algorithm, requires less time during classification. In KNN, the training dataset is directly stored in memory and used for making predictions. The algorithm identifies the k closest objects in the data that is trained to the input instance by employing an Euclidean distance metric. The class with the highest number of votes among these nearby classes is assigned to categorize the unknown data. The effectiveness of the classifier hinges on choosing the optimal value for k. In this study, the KNN classifier explores odd values of k up to 25. Distance metrics, such as Euclidean and Manhattan, can be used to determine the proximity of the new instance to the training examples, with Euclidean distance being the most commonly used. In this work, Euclidean distance is used for the KNN classification algorithm. The following specific equation can be used to retrieve the Euclidean distance (d) from one point to another point (y, z), and

$$d(y, z) = \sqrt{\sum_{i=1}^{n} (y_i - z_i)^2}. \tag{5}$$

6) MULTI-LAYER PERCEPTRON (MLP)

The perceptron algorithm [47] uses a set of training examples to train a prediction vector that correctly predicts the output labels for the input instances in the training set. It achieves this by updating the network weights through a quadratic programming problem with linear constraints in stead of a nonconvex, unconstrained optimization problem used in traditional neural network training. The algorithm runs over the training set until a correct prediction vector is found. In the test set, the labels of the cases are predicted by considering the prediction vector obtained.

7) SUPPORT VECTOR MACHINE (SVM)

The Support Vector Machine (SVM) [40], [41] is frequently utilized as a supervised learning method for both regression and classification analysis. These models utilize specific learning algorithms and are capable of handling both linear and non-linear classification tasks by employing the "kernel trick" that maps the inputs to higher-dimensional feature spaces. By maximizing the margin, which refers to the distance between classes, SVMs aim to minimize classification errors.

8) K-MEANS

The k-means algorithm is a straightforward unsupervised learning method used to solve the clustering problem [48], [49]. The data set is categorized into a pre-determined number of clusters (k) without using labeled data. It is a method for transforming general guidelines into highly accurate prediction rules by consistently discovering classifiers that perform slightly better than random.

Subsequently, we expand the centrality measure by utilizing the relative change in centrality.

IV. GENERALIZING THE RELATIVE CHANGE IN LOCAL CENTRALITY

In this section, we present a generalized centrality measure that examines the relative variation in centrality. By evaluating the influence of centrality measures when removing a vertex, we extend the notion of assessing centrality for nodes within a network. While existing centrality measures focus on the network's local and global structures, our approach is tailored to local centrality measures, analyzing the impact of node removal on the relative change in various centrality measures.

Here, we introduce a local measure that relies on the graph's local structure, requiring access to local network data. Let us consider the neighborhood level denoted as L for a given vertex v within the graph G. The L may consider values from 0 up to the graph's diameter. Consider the following definition for the local measure:

$$LRAvg\mathscr{C}_L(v) = \frac{\left|Avg\mathscr{C}[G_{N_L(v)} \setminus v] - Avg\mathscr{C}[G_{N_L(v)}]\right|}{Avg\mathscr{C}[G_{N_L(v)}]}. \tag{6}$$

The graph $G_{N_L(v)} \setminus v$ represents the outcome of eliminating vertex v from the graph $G_{N_L(v)}$. Determine the $Avg\mathscr{C}$ associated for induced subgraph $G_{N_L}(v)$ in the graph G, where $G_{N_L}(v)$ consists of the vertices in the neighborhood $N_L(v)$. The centrality measure $LRAvg\mathscr{C}$ is a local method

used to calculate centrality for a vertex. It considers only the neighboring vertices within a certain level L (see for all notations in table 2).

When the centrality measure $\mathscr{C}$ is treated as degree (D) [50], then average degree is computed as

$$AD[G] = \frac{\sum_{v \in V} d_v}{N}, \tag{7}$$

where d_v indicates the vertex degree. The average degree [35] of local relative change for node v in G evaluate as

$$LRAD_L(v) = \frac{\left|AD[G_{N_L(v)} \setminus v] - AD[G_{N_L(v)}]\right|}{AD[G_{N_L(v)}]}. \tag{8}$$

When the centrality measure $\mathscr{C}$ is considered as the clustering coefficient (CC) [51], the average CC is computed as

$$ACC[G] = \frac{\sum_{v \in V} CC(v)}{N}, \tag{9}$$

where $CC(v)$ denotes the vertex clustering coefficient.

The average clustering coefficient [35] can be determined using the vertex v of local relative change, which is determined as

$$LRACC_L(v) = \frac{\left|ACC[G_{N_L(v)} \setminus v] - ACC[G_{N_L(v)}]\right|}{ACC[G_{N_L(v)}]}. \tag{10}$$

When the centrality measure $\mathscr{C}$ is considered to be as Katz centrality (K) [52], then the average Katz centrality (AK) is calculated as

$$AK[G] = \frac{\sum_{v \in V} K(v)}{N}, \tag{11}$$

where $K(v)$ represents the vertex Katz centrality. The local relative change in average Katz centrality [35] is defined as

$$LRAK_L(v) = \frac{\left|AK[G_{N_L(v)} \setminus v] - AK[G_{N_L(v)}]\right|}{AK[G_{N_L(v)}]}. \tag{12}$$

We opted to emphasize classical centrality measures, along with LR measures, aligning with our study's particular context and objectives. Due to the high computation time associated with individual classical centrality measures, we employed neighborhood level (L) values adjustable within the range of 1 to the diameter. Through examinations, we determined that setting L to half of the diameter effectively captures more local neighbor's information while maintaining lower time complexity. Some studies have looked into the rich-club [34] and directive properties of complex networks to identify exact nodes in the network and their influence [53]. Additionally, some studies use machine learning approaches to determine influential nodes in their particular contexts, including user influence over Facebook [33]. However, the universal propagation scenario is not examined in these works.

V. IMPLEMENTATION

A methodology for using machine learning methods is used to recognize and relate the nodes in the network that are highly influential, as shown in Figure 1. The key step in constructing a training dataset for model development involves highlighting the dependency of each node's feature vector on conventional centrality measures and the infection probability beta within the model's propagation. After constructing this dataset, the spreading ability of each person can be determined by employing the SIR model and IC diffusion models with the infection probability rate, which acts as the post-computational label for the individual. Subsequently, diverse ML approaches are applied to the training data to uncover categorization patterns and generate corresponding classification models. Leveraging these trained models, assessing the significance of any person within the same network or another node from various networks becomes possible. A framework model follows five steps: (1) Evaluating node influence based on the SIR and IC model. (2) Computing the label of the node based on its influence. (3) Choose the features of the node. (4) Extracting the real label of node based on SIR and IC models. (5) Node is labeled based on centrality methods. These five steps are explained more in detail in the following subsections (see Figure 1).

A. ASSESSING NODE IMPACT USING SIR AND IC MODELS

In this study, we utilize the Susceptible-Infectious-Recovered (SIR) model widely recognized to evaluate the spread of nodes in a network and propose a node labeling strategy based on their propagation ability. The SIR, prevalent epidemic model classifies nodes as susceptible, infected, or recovered. Infectious nodes can transmit the infection to susceptible neighbors through the infection and recovery rate, eventually leading to recovery. Recovered nodes become immune to further infections and cannot infect others. The dynamics of the SIR model are characterized by randomness and interactions between nodes. Our objective is to identify an initially infected node while considering the magnitude of the resulting epidemic as a measure of its true influence. The extent of the epidemic in our research is determined by the total infected and recovered nodes, with nodes contributing to a larger outbreak considered to have a greater impact. For our results, we considered the recovery rate λ to a fixed value of 1. As for the infection rate β, we varied it within the range of 0.1 and 0.2.

B. ASSIGNING LABELS BASED ON NODE'S INFLUENCE

Our objective is to identify significant nodes within networks using machine learning techniques that involve labeling. However, directly assigning labels to nodes based on the exact scale of outbreaks obtained from SIR tests can result in excessive granularity. Therefore, in this study, we establish the labels based on the following concept that

$$label_i = \left[\frac{s_i - \text{MinScale}}{range}\right] + 1. \tag{13}$$

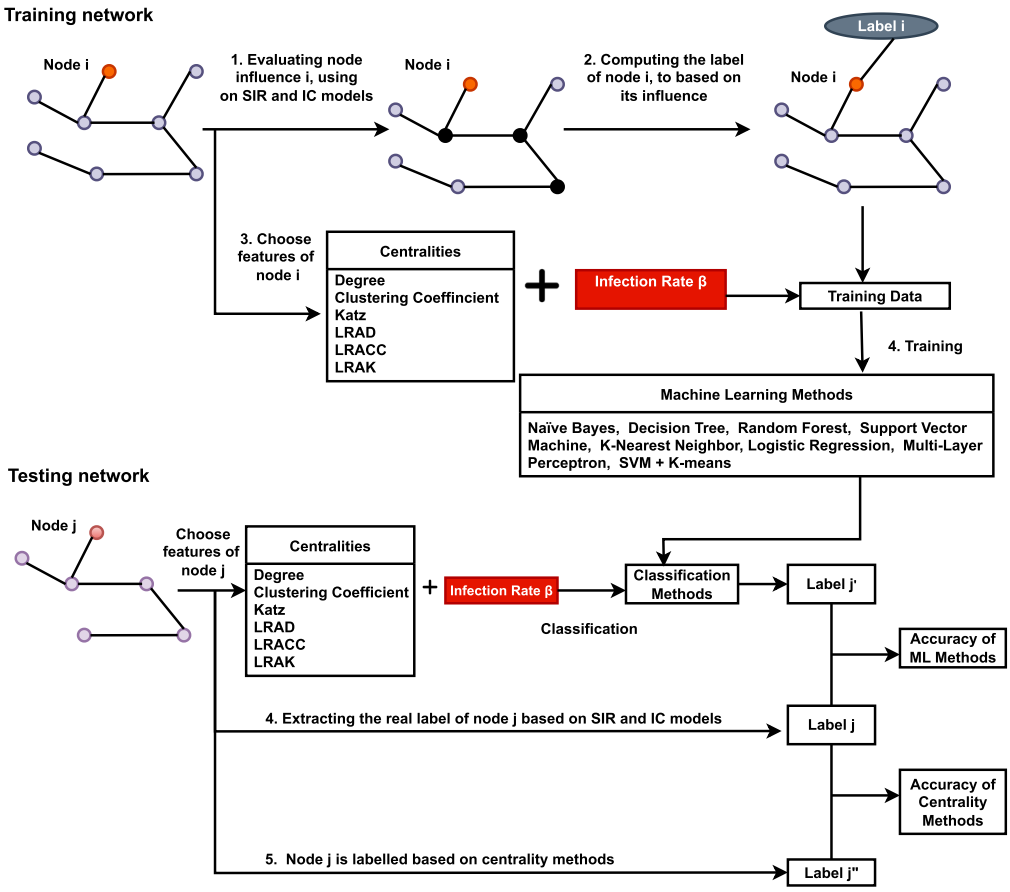

FIGURE 1. An overview of the machine learning-based methodology for node influence identification in networks. This figure illustrates the five-step process: (1) influence evaluation of nodes using the SIR and IC models, (2) computation of node labels based on influence, (3) selection of node features, (4) extraction of real node labels using SIR and IC models, and (5) labeling of nodes based on centrality methods. The methodology integrates conventional centrality measures, infection probability beta, and machine learning techniques to effectively determine and categorize the spreading ability of nodes within a network.

In this context, MinScale denotes the smallest observed scale of outbreaks in the SIR results, while S_i represents the final scale of an outbreak associated with node i. The range indicates the interval size between labels, as defined in Equation (2), and is given by

$$range = \frac{\text{MaxScale} - \text{MinScale}}{N}. \tag{14}$$

In this case, MaxScale signifies the highest scale attained by the outbreaks, while N denotes the number of labels present. Specifically, if a node's computation result is $N + 1$, the node will be assigned the label of N.

Labeling the vertices in the network is important as we try to identify the node's influence by utilizing machine-learning techniques. SIR experiments produce too fine a granularity when utilizing an outbreak final scale. Where N is the number of labels, MinScale denotes the outbreak's minimum final scale, and MaxScale denotes the outbreak's maximum final scale, as determined by SIR and IC methods.

Assigning labels to the vertices in the network is essential as our goal is to determine the influence of each node using machine learning techniques. However, applying the precise outbreak scale discovered through SIR research leads to an inappropriate level of granularity.

To overcome this challenge, we introduce the following variables Minscale, Maxscale, and N, where N is the number of labels. The outbreak's minimum and maximum final scale is determined by SIR and IC methods. By taking these variables into account, we can establish a more appropriate and manageable granularity level when labeling the network's vertices.

C. DEFINING NODE'S ATTRIBUTES

The key objective in creating the training data set is to establish the features and their respective values for each node, as these features play a fundamental role in evaluating the significance of nodes. Previous studies have proposed diverse centrality measures to evaluate node importance. Therefore, we can select specific traditional centrality measures as features in our dataset. Furthermore, considering that the scales of node outbreaks can vary with different infection

rates, the infection rate becomes a critical factor that affects node propagation ability. Thus, in this study, the infection rate is included as one of the node attributes. To capture a comprehensive range of structural aspects of nodes, our study incorporates multiple types of centrality approaches. By incorporating various centrality measures, we aim to encompass different perspectives on node importance within the data set.

In addition, this study applies normalization to the centrality features to prevent overfitting and enhance the generalizability of the trained classifier to different networks. The normalization technique employed for each centrality feature, denoted as k, is as follows

$$f_k = \frac{P_k}{N}, \tag{15}$$

where P_k indicates the ranking position based on centrality value k, and N suggests the nodes exist in the network.

1) TRAINING ML MODELS

After constructing the training data, we implemented various machine learning methodologies to build ML models. In our study, we consider the following eight commonly used machine learning algorithms: 1) NBA, 2) DT, 3) RF, 4) SVM, 5) KNN, 6) LR, 7) MLP, and 8) SVM+k-means. To create our machine learning models, we utilize the scikit-learn package. We use grid search to fine-tune the pertinent hyperparameters in order to improve the method's performance.

2) NODES LABELING BASED ON CENTRALITY MEASURES

To evaluate and know the effectiveness of the ML model that is shown in Figure 1, we introduce six centrality methods for comparison: 1) D, 2) CC, 3) K, 4) LRAD, 5) LRACC, and 6) LRAK. The experimentation begins by utilizing a real-world network for both training and testing purposes. The nodes are randomly split between 30% for testing and 70% for training data. Subsequently, machine learning methods are applied during the testing phase to generate classification outputs. By comparing these outputs with the actual labels, we calculate the accuracy of the ML methods. Furthermore, we compare the accuracy achieved by ML methods with that of traditional centrality methods.

It is important that the centrality methods and infection rate serve as labels for the nodes in this process. The results demonstrate that machine learning methods exhibit higher accuracy than traditional centrality methods. For a specific label i, the calculation of the distribution's location is $\sum_{j=i+1}^{N} P_j \sum_{j=i}^{N} P_j$, where P_j represents the label proportion.

VI. DATASETS

To evaluate and observe the efficacy of our approach, this study incorporates various traditional information transmission scenarios, namely USAir97, bio-celegans, ca-netscience, and web-polblogs. These scenarios were carefully selected to encompass a diverse range of network characteristics. The datasets utilized for these networks were obtained from a reliable source [54]. The USAir97 network dataset represents the connectivity between airports in the United States during 1997. Widely recognized as a benchmark dataset in network analysis and research, it captures the relationships between airports based on direct flight connections. In USAir97, each node represents an airport, and the edges depict direct flights between airports. The USAir97 network consists of 332 nodes and 2126 edges. On the other hand, the bio-celegans network dataset provides a representation of the neural connectivity in the nematode worm Caenorhabditis elegans. This dataset illustrates the connections between various neurons within the worm's nervous system, where neurons are represented as nodes and synapses as edges.

The bio-celegans contain 453 nodes and 2025 edges. The ca-netscience network dataset illustrates a collaboration network among authors in network science. In this network, authors are represented as nodes, and the edges depict co-authorship relationships. The ca-netscience network consists of 379 authors and 914 edges. The web-polblogs network dataset portrays the political blogosphere during the 2004 US presidential election. Nodes in this network represent political blogs, while the edges represent hyperlinks between them. The web-polblogs network encompasses 643 blogs and 2280 hyperlinks between them. These carefully selected networks provide a solid foundation for evaluating the effectiveness of our methodology in various network analysis scenarios.

VII. RESULTS AND ANALYSIS

Our study involved several experiments aimed at comparing the effectiveness of an ML model with traditional centrality methods across various scenarios. In this research study, we conducted experiments on a high-performance computing system boasting 128GB of RAM. The system operated on the Windows 11 Pro operating system, featuring an 11th Gen Intel® Core i9-1190 processor with 3.8 GHz speed. The system type was a 64-bit operating system, and Python 3.7.10 was employed for data analysis. Implementation of this work in Python code is available in GitHub link in [55]. Graphical representations were crafted using Origin PRO software. The experiments were divided into two sections. In the first section, we focused on evaluating the performance of the ML model by training and testing it on data within the same network. The second section involved examining scenarios where the training and testing nodes belonged to different networks, allowing us to assess the scalability of the ML model. One notable observation was the significant impact of the number of labels, which represent node categories, on the classification accuracy. As the number of labels increased, the accuracy of categorization decreased. To further investigate this effect, we trained seven classifiers using seven algorithms for each label count and evaluated their classification performance. In addition to label count, we also considered the influence of the infection rate on the spreading ability within the training phase. Our approach

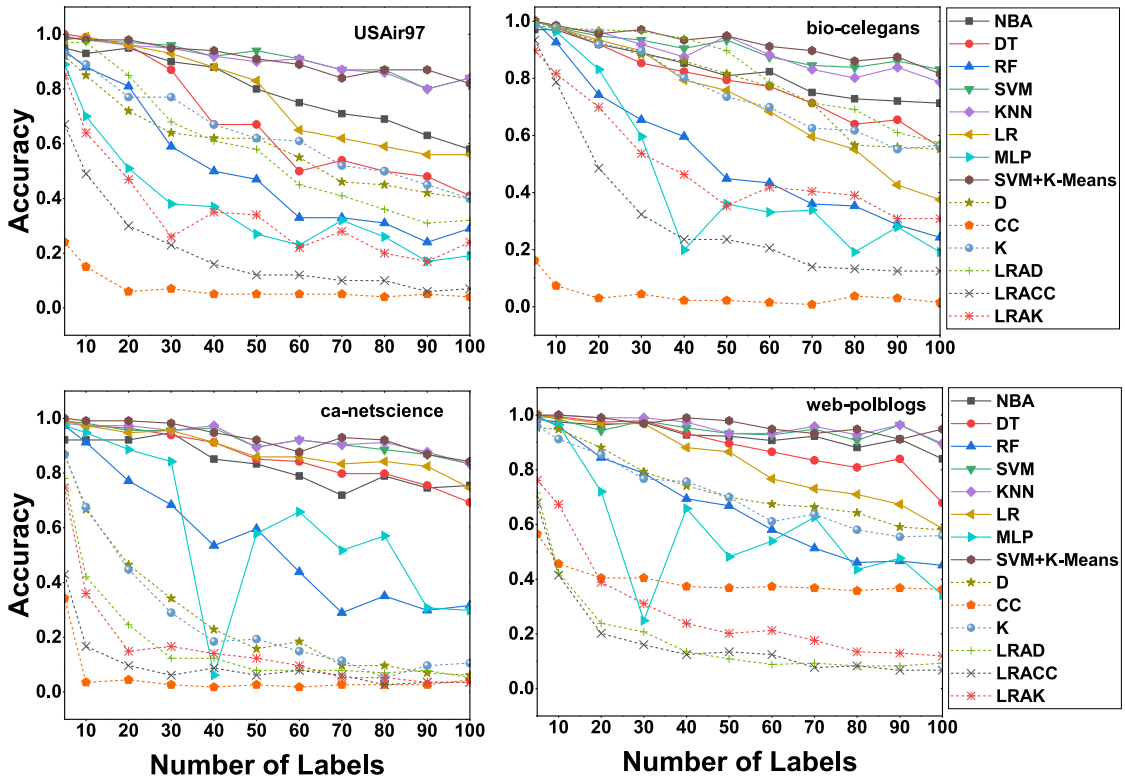

FIGURE 2. ML algorithms are applied to various data sets for training and testing to determine the accuracy of proposed centralities (0.1).

involved incorporating various data points, each representing different infection rates and label values for every node. This enabled the ML algorithms to learn about the effect of infection rate on the significance of nodes. For our experiments, we selected infection rates ranging from 0.1 to 0.2. During the testing phase, we simplified the studies by using two infection rates (0.1 and 0.2) to predict the performance of the classifiers.

A. TESTING AND TRAINING CONDUCTED ON THE SAME NETWORK

The tests were split into two parts. We used the same network in the first section for both training and testing. In the second portion, we tested with a different network from the one we trained with. The quantity of labels (number of node categories) has a significant impact on categorization accuracy. The accuracy of categorization decreases as the number of labels increases.

Each network's nodes were randomly divided into 30% and 70% for testing and training data. The results, considering various infection rates and label counts, are presented. The overall findings indicate that machine learning methods outperformed centrality approaches such as degree (D), clustering coefficient (CC), Katz (K), LRAD, LRACC, and LRAK. These results are visualized in Figure 2, and Figure 3. Moreover, it was observed that the infection rate, as well as the number of labels, influenced the accuracy of ML techniques. Across all networks, accuracy decreased as the number of labels increased and slightly decreased with higher infection rates. This decrease in accuracy can be attributed to the reduced distinguishability of node spreading capabilities as infection rates rise, leading to compromised classification performance. Specifically, the SVM+k-means algorithm demonstrated superior accuracy compared to other ML models and centrality approaches. The SVM and KNN models also achieved high accuracy to other machine learning models. For small-sized networks, the accuracy of machine learning models initially increased and then decreased. However, in large networks, accuracy either remained constant or decreased as the number of labels increased. In contrast, classical centrality approaches exhibited an initial increase in accuracy for a smaller number of labels, followed by stabilization as the number of labels increased at a constant infection rate.

In the USAir97 network, the SVM+k-means method demonstrated higher accuracy than other D, CC, K, LRAD, LRACC, and LRAK centrality approaches, as shown in Figure 2, and Figure 3. Subsequently, the KNN method achieved higher accuracy compared to other centrality measures. Similarly, in the bio-celegans network, the SVM+k-means method yielded higher accuracy than other centrality approaches, as depicted in Figure 2, and Figure 3. Later, the SVM method exhibited better accuracy than the other methods. The SVM+k-means technique achieved good

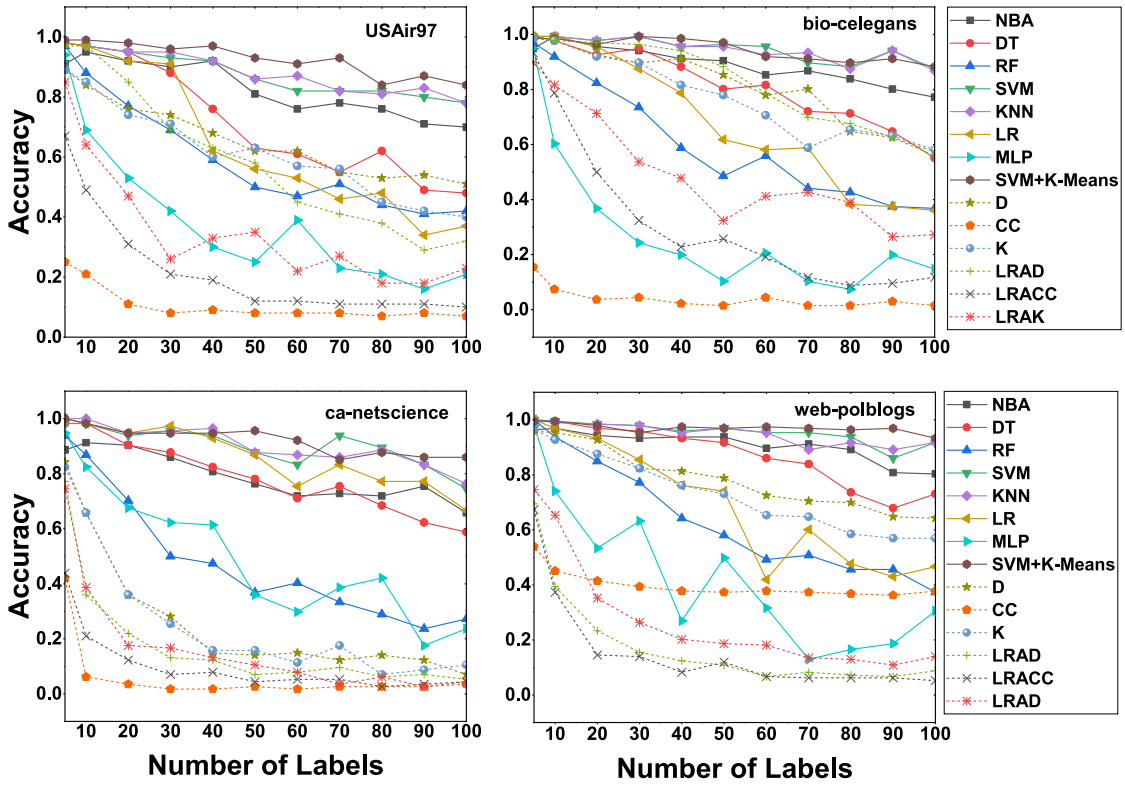

FIGURE 3. ML algorithms are applied to various data sets for training and testing to determine the accuracy of proposed centralities (0.2).

accuracy in the ca-netscience network. Next, KNN method is performed with higher accuracy than DT, RF, SVM, LR, MLP, D, CC, K, LRAD, LRACC, and LRAK, as shown in Figure 2, and Figure 3. In the web-polblogs dataset, the SVM+k-means method demonstrated higher accuracy than other methods as shown in Figure 2 and Figure 3.

B. TESTING AND TRAINING ON VARIOUS NETWORKS

In this section, our experiments aim to assess the generalizability of the proposed model by evaluating the performance of ML models when training and test data are obtained from dissimilar networks. We consider the impact of label quantity and infection rate on the accuracy of each model. The experimental results consistently demonstrate that irrespective of the approach employed, classification accuracy tends to decrease as the number of labels and infection rate increase. Notably, regardless of network size, machine learning approaches exhibit significantly lower classification accuracy than centrality methods. These results are depicted in Figure 4, Figure 5, Figure 6, Figure 7, Figure 8, and Figure 9.

When comparing the LRAD centrality method with machine learning models and other centrality approaches such as D, CC, K, LRACC, and LRAK, we observe that it achieves the highest accuracy. Additionally, the accuracy of the LRACC and K centrality methods remains stable across various infection rates. In contrast, the accuracy of machine learning classifiers displays multiple fluctuations, initially increasing, then decreasing, and eventually stabilizing as the infection rate, as well as the number of labels, increase. In our analysis of the USAir97 and bio-celegans networks, we found that the LRAD and D centrality methods demonstrated superior performance compared to machine learning approaches, as illustrated in Figure 4. The conventional centrality measures achieved higher accuracy than the machine learning methods in these networks. To evaluate their performance, we used USAir97 as the training set and bio-celegans as the testing set, as well as vice versa. Additionally, we employed ca-netscience as the training set and web-polblogs as the testing set, as depicted in Figure 5. Within the ca-netscience network, the D and K centrality methods outperformed the machine learning approaches.

In the evaluation of the USAir97 and web-polblogs networks, we observed that the degree and Katz centrality methods exhibited superior performance compared to machine learning approaches, as presented in Figure 6. Traditional centrality measures achieved higher accuracy than the machine learning methods. For these networks, we utilized USAir97 as the training set and web-polblogs as the testing set, and vice versa. Similarly, in the case of the ca-netscience and bio-celegans networks, we found that the LRAK and K centrality methods performed well. We employed ca-netscience as the training set and bio-

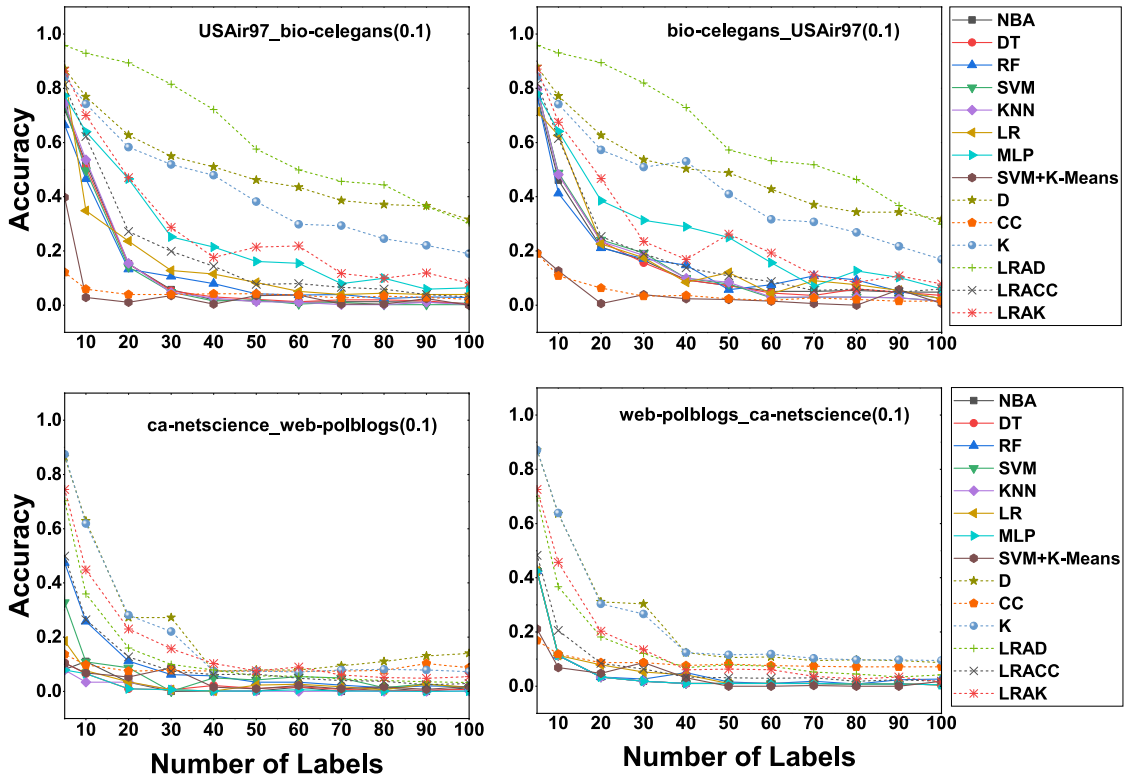

FIGURE 4. Application of machine learning algorithms to diverse data sets for evaluating the accuracy of proposed centralities (0.1).

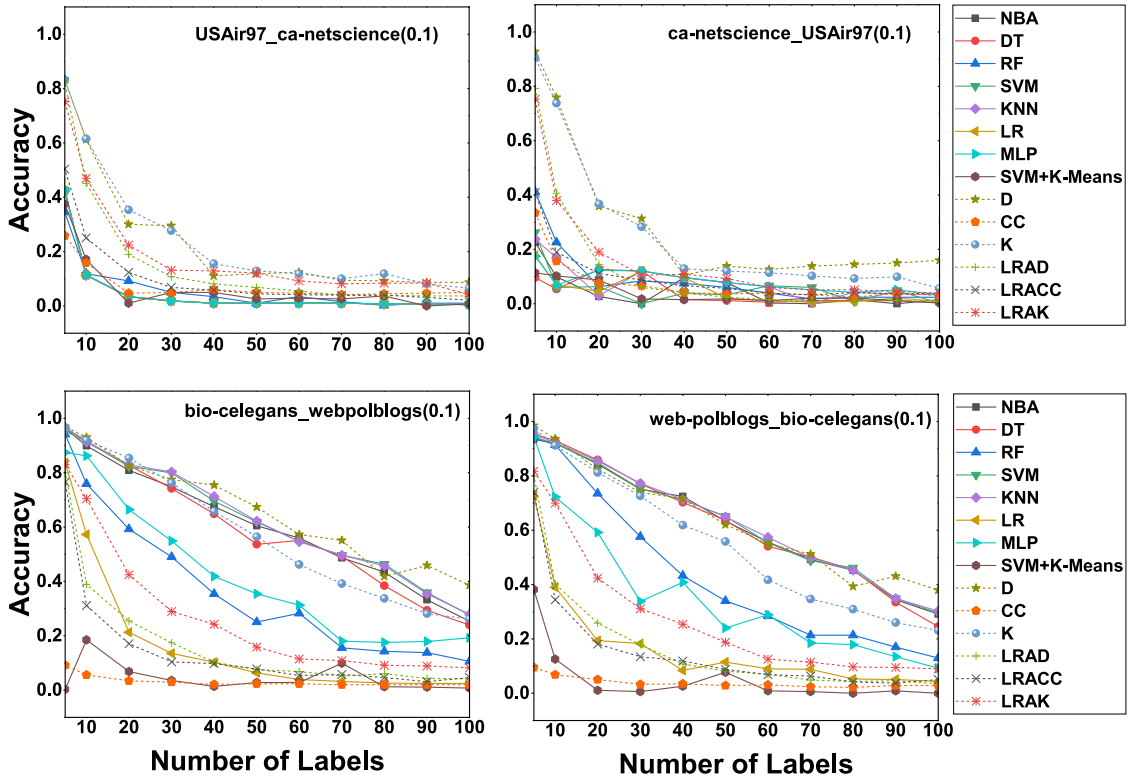

FIGURE 5. Experimental validation of proposed centralities through machine learning algorithms on diverse data sets (0.1).

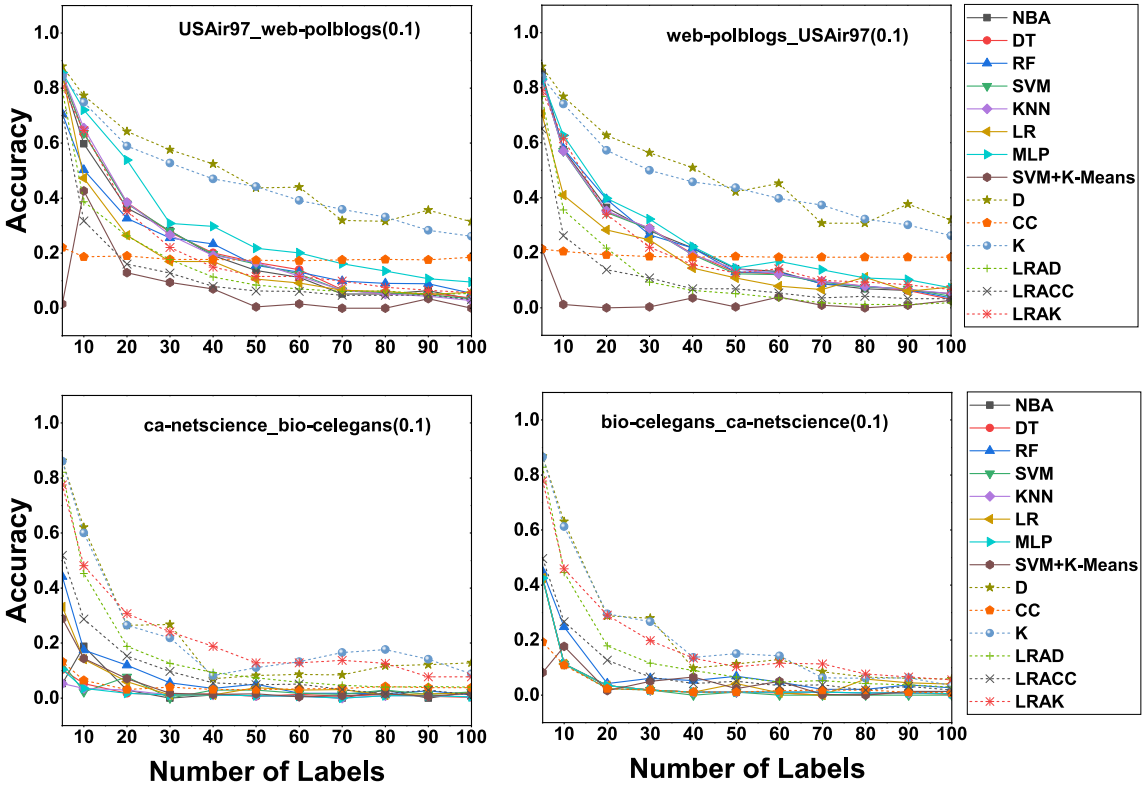

FIGURE 6. Evaluation of centralities accuracy utilizing machine learning algorithms on varied data sources (0.1).

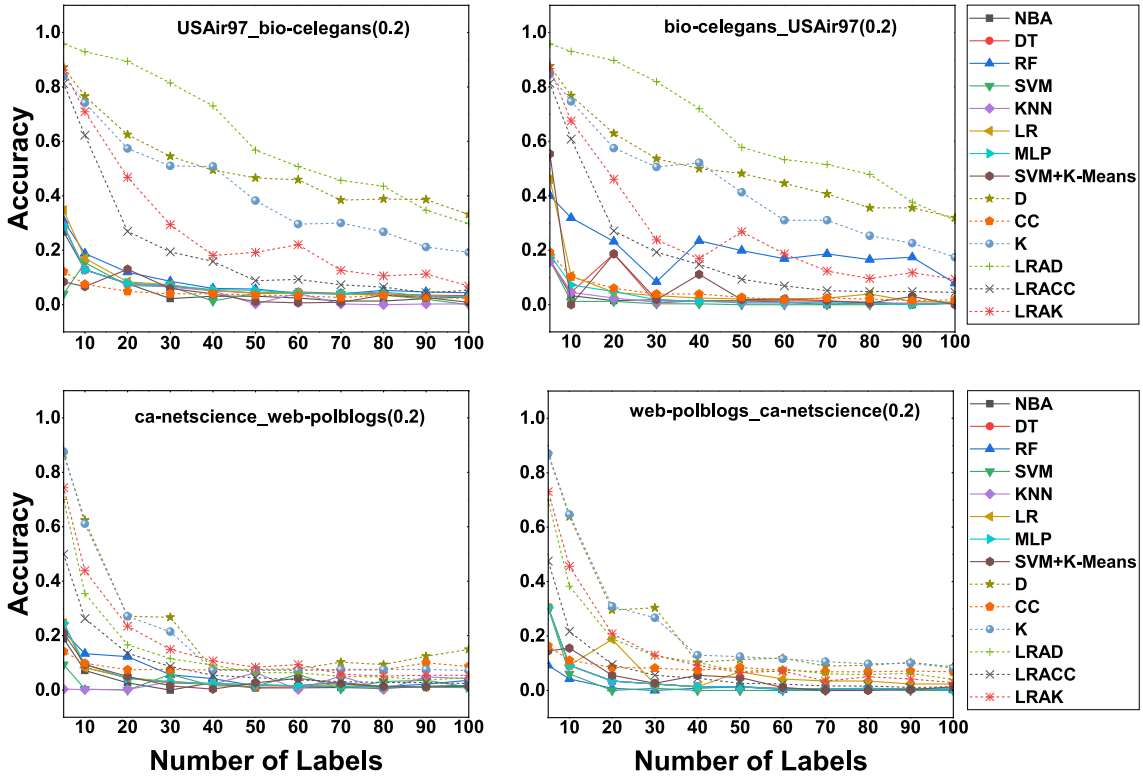

FIGURE 7. Machine learning algorithms are applied to various data sets for training and testing to determine the accuracy of proposed centralities (0.2).

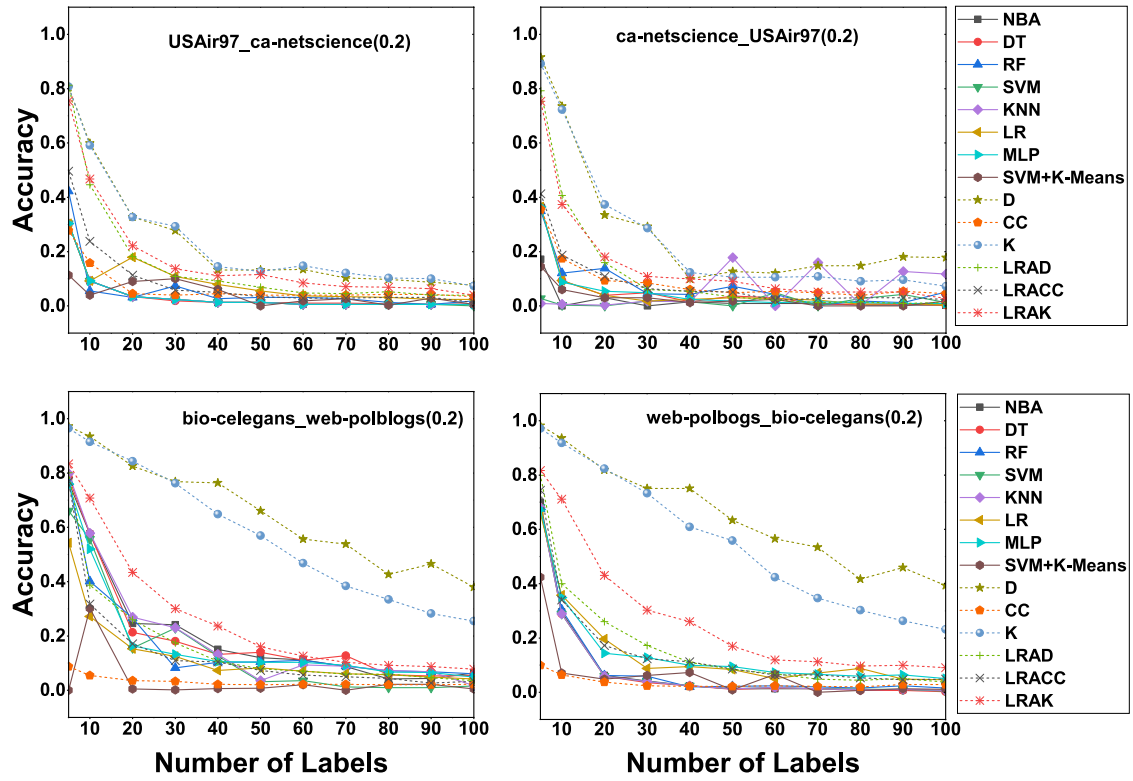

FIGURE 8. Machine learning algorithms employed for training and testing across diverse data sets to assess proposed centralities accuracy (0.2).

celegans as the testing set, and vice versa, as illustrated in Figure 6.

In Figure 7, we can notice that the LRAD and degree centrality methods outperformed machine learning approaches in the USAir97 and bio-celegans networks. Traditional centrality measures demonstrated higher accuracy compared to machine learning methods. For the training and testing sets, we utilized the USAir97 network and bio-celegans, respectively, and vice versa. Furthermore, in the ca-netscience and web-polblogs networks, we employed ca-netscience as the training set and web-polblogs as the testing set, as depicted in Figure 7. In both networks, the degree and Katz centralities exhibited superior accuracy compared to other centrality measures.

In Figure 8, we notice that the degree centrality method outperformed machine learning approaches in the USAir97 and ca-netscience networks. Conventional centrality measures achieved higher accuracy compared to machine learning methods. For the training and testing sets, we used USAir97 and ca-netscience, respectively, and vice versa. Similarly, in the case of the bio-celegans and web-polblogs networks, we found that the degree centrality method exhibited good performance. We employed bio-celegans as the training set and web-polblogs as the testing set, and vice versa, as depicted in Figure 8.

In Figure 9, we can notice that the degree and Katz centrality methods demonstrated superior performance over machine learning approaches in the USAir97 and web-polblogs networks. Traditional centrality measures achieved higher accuracy compared to ML methods. The USAir97 network was used as the training set, while web-polblogs served as the testing set, and vice versa. Similarly, in the case of the bio-celegans and ca-netscience networks, we employed bio-celegans as the training set, while ca-netscience served as the testing set, as depicted in Figure 9.

C. DISCUSSIONS

To pinpoint crucial nodes within intricate networks, conventional centrality methods are typically constructed by directly analyzing specific topological structures of the network. However, this approach imposes constraints on both performance and adaptability when identifying influential nodes in propagation scenarios. To address these limitations, this study reframed the challenge of identifying influential nodes as a classification problem. The provided model utilizes machine learning techniques to improve the precision and adaptability of identifying influential nodes. The experimental results presented in this section highlight the substantial impact of network scale on the effectiveness of the introduced machine learning model, especially when training and testing networks exhibit variations. The findings suggest that the machine learning classifier demonstrates superior performance compared to centrality methods when trained

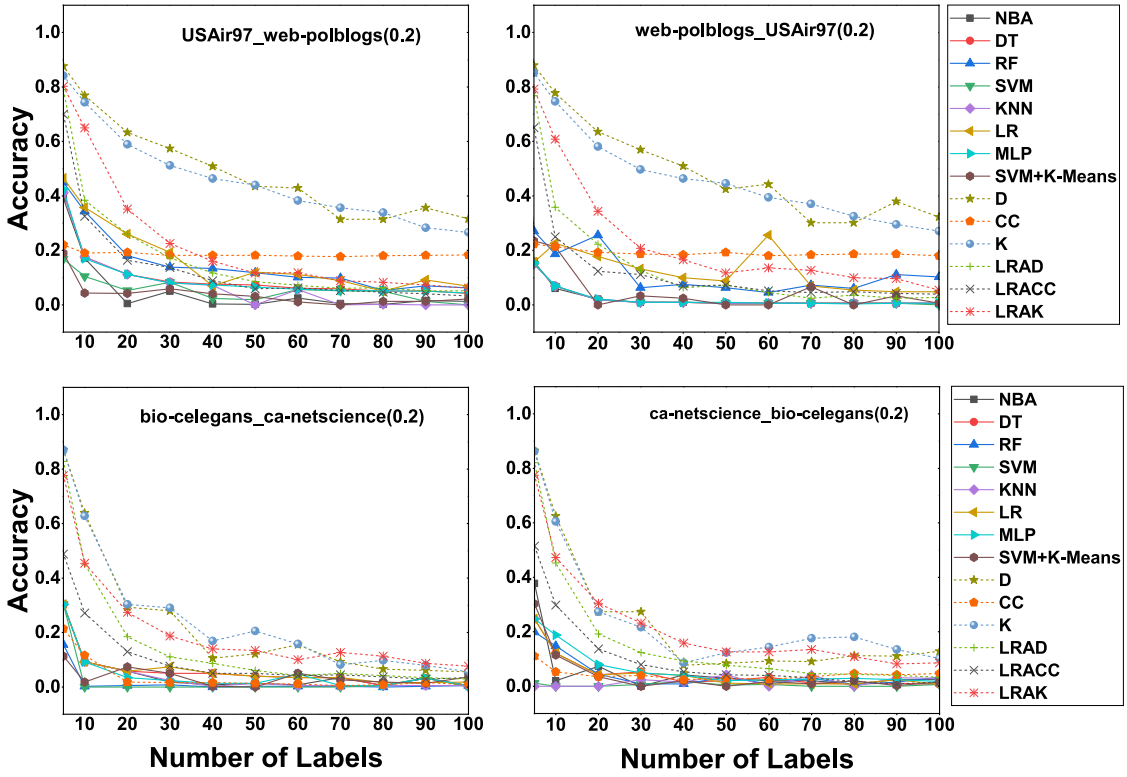

FIGURE 9. Utilizing machine learning algorithms to gauge the accuracy of proposed centralities through training and testing on diverse data sets (0.2).

on a small-scale network and later tested on a large-scale network.

VIII. CONCLUSION

In our study, we introduced three innovative centrality measures that leverage local structural details. These novel metrics, LRACC, LRAD, and LRAK, were developed by considering the relative changes in CC, degree, and Katz centrality following node deletion. Our analysis revealed that these proposed centralities exhibited superior performance compared to traditional metrics like D, CC, and K, as well as machine learning algorithms. Centrality methods such as degree, clustering coefficient, Katz, LRAD, LRACC, and LRAK are primarily designed to identify important nodes to spread the disease in complex networks by focusing on their topological structures. To address centrality measures that generalize based on the relative alterations in CC, D, and K centrality after the removal of nodes. This study approaches the problem of identifying influential nodes as a classification problem and proposes the utilization of machine learning models for this purpose. Unfortunately, there are only a few studies that have explored the application of machine learning techniques in the specific context of discovering vital nodes.

By conducting SIR and independent cascade experiments, the study suggests a person labeling method and employs machine learning algorithms to refine the classification criteria based on node attributes and labeling results. The proposed SVM+k-means algorithm exhibits higher accuracy compared to conventional centrality methods when trained and tested within the same network. However, when applied to a different network for training and testing, the betweenness centrality method demonstrates superior accuracy compared to machine learning approaches. Notably, when the machine learning classifiers are trained in a larger network and then tested in a smaller network, they perform less effectively than centrality methods. While machine learning classifiers are more efficient than centrality methods for training and testing in a larger network, their performance deteriorates when tested in a smaller network after being trained in a larger network. Exploring novel avenues for research, a promising direction involves devising centrality measures that optimize information dissemination efficiently. One approach worth considering is the integration of both local and global centrality metrics, aiming to maximize information spread while minimizing computational time. Additionally, it is valuable to explore the relative changes observed in existing centralities within the literature, thereby enhancing our understanding of the effectiveness of current methods or deep learning methods.

REFERENCES

[1] G. Pallis, D. Zeinalipour-Yazti, and M. D. Dikaiakos, "Online social networks: Status and trends," in *New Directions in Web Data Management 1*. Berlin, Germany: Springer, 2011, pp. 213–234.

[2] J. Heidemann, M. Klier, and F. Probst, "Online social networks: A survey of a global phenomenon," *Comput. Netw.*, vol. 56, no. 18, pp. 3866–3878, Dec. 2012.
[3] L. C. Freeman, "Centrality in social networks conceptual clarification," *Social Netw.*, vol. 1, no. 3, pp. 215–239, Jan. 1978.
[4] U. Brandes, "A faster algorithm for betweenness centrality," *J. Math. Sociol.*, vol. 25, no. 2, pp. 163–177, 2001.
[5] K. Okamoto, W. Chen, and X.-Y. Li, "Ranking of closeness centrality for large-scale social networks," in *Proc. Int. Workshop Frontiers Algorithmics*. Cham, Switzerland: Springer, 2008, pp. 186–195.
[6] P. Bonacich, "Some unique properties of eigenvector centrality," *Social Netw.*, vol. 29, no. 4, pp. 555–564, Oct. 2007.
[7] P. Bonacich, "Factoring and weighting approaches to status scores and clique identification," *J. Math. Sociol.*, vol. 2, no. 1, pp. 113–120, Jan. 1972.
[8] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. A. Makse, "Identification of influential spreaders in complex networks," *Nature Phys.*, vol. 6, no. 11, pp. 888–893, Nov. 2010.
[9] S. X. Zhao, R. Rousseau, and F. Y. Ye, "H-degree as a basic measure in weighted networks," *J. Informetrics*, vol. 5, no. 4, pp. 668–677, Oct. 2011.
[10] P. Hage and F. Harary, "Eccentricity and centrality in networks," *Social Netw.*, vol. 17, no. 1, pp. 57–63, Jan. 1995.
[11] L. Katz, "A new status index derived from sociometric analysis," *Psychometrika*, vol. 18, no. 1, pp. 39–43, Mar. 1953.
[12] E. Estrada and J. A. Rodríguez-Velázquez, "Subgraph centrality in complex networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 71, no. 5, May 2005, Art. no. 056103.
[13] J. Ding and A. Zhou, "Eigenvalues of rank-one updated matrices with some applications," *Appl. Math. Lett.*, vol. 20, no. 12, pp. 1223–1226, Dec. 2007.
[14] K. Das, S. Samanta, and M. Pal, "Study on centrality measures in social networks: A survey," *Social Netw. Anal. Mining*, vol. 8, no. 1, pp. 1–11, Dec. 2018.
[15] S. Pei and H. A. Makse, "Spreading dynamics in complex networks," *J. Stat. Mech., Theory Exp.*, vol. 2013, no. 12, Dec. 2013, Art. no. P12002.
[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[17] Y. Wang, Z. Wei, and J. Cao, "Epidemic dynamics of influenza-like diseases spreading in complex networks," *Nonlinear Dyn.*, vol. 101, no. 3, pp. 1801–1820, Aug. 2020.
[18] M. K. Enduri and S. Jolad, "Dynamics of dengue disease with human and vector mobility," *Spatial Spatio-Temporal Epidemiol.*, vol. 25, pp. 57–66, Jun. 2018.
[19] Y. Zhang, S. Wang, P. Phillips, and G. Ji, "Binary PSO with mutation operator for feature selection using decision tree applied to spam detection," *Knowl.-Based Syst.*, vol. 64, pp. 22–31, Jul. 2014.
[20] C. Guo, L. Yang, X. Chen, D. Chen, H. Gao, and J. Ma, "Influential nodes identification in complex networks via information entropy," *Entropy*, vol. 22, no. 2, p. 242, Feb. 2020.
[21] K. Hajarathaiah, M. K. Enduri, S. Anamalamudi, T. S. Reddy, and S. Tokala, "Computing influential nodes using the nearest neighborhood trust value and PageRank in complex networks," *Entropy*, vol. 24, no. 5, p. 704, May 2022.
[22] K. Hajarathaiah, M. K. Enduri, S. Anamalamudi, and A. R. Sangi, "Algorithms for finding influential people with mixed centrality in social networks," *Arabian J. Sci. Eng.*, vol. 48, no. 8, pp. 10417–10428, Aug. 2023.
[23] K. Berahmand, A. Bouyer, and N. Samadi, "A new centrality measure based on the negative and positive effects of clustering coefficient for identifying influential spreaders in complex networks," *Chaos, Solitons Fractals*, vol. 110, pp. 41–54, May 2018.
[24] Z. Lv, N. Zhao, F. Xiong, and N. Chen, "A novel measure of identifying influential nodes in complex networks," *Phys. A, Stat. Mech. Appl.*, vol. 523, pp. 488–497, Jun. 2019.
[25] K. Hajarathaiah, M. K. Enduri, and S. Anamalamudi, "Efficient algorithm for finding the influential nodes using local relative change of average shortest path," *Phys. A, Stat. Mech. Appl.*, vol. 591, Apr. 2022, Art. no. 126708.
[26] G. Zhao, P. Jia, C. Huang, A. Zhou, and Y. Fang, "A machine learning based framework for identifying influential nodes in complex networks," *IEEE Access*, vol. 8, pp. 65462–65471, 2020.
[27] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," *BMC Med. Informat. Decis. Making*, vol. 19, no. 1, pp. 1–16, Dec. 2019.
[28] A. A. Rezaei, J. Munoz, M. Jalili, and H. Khayyam, "Vital node identification in complex networks using a machine learning-based approach," 2022, *arXiv:2202.06229*.
[29] T. Opsahl, "Degree centrality in a weighted network," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 77, no. 4, 2008.
[30] L. Lü, Y.-C. Zhang, C. H. Yeung, and T. Zhou, "Leaders in social networks, the delicious case," *PLoS ONE*, vol. 6, no. 6, Jun. 2011, Art. no. e21202.
[31] P. Li, Y. Ren, and Y. Xi, "An importance measure of actors (set) within a network," *Syst. Eng.*, vol. 22, no. 4, pp. 13–20, 2004.
[32] C. Dangalchev, "Residual closeness in networks," *Phys. A, Stat. Mech. Appl.*, vol. 365, no. 2, pp. 556–564, Jun. 2006.
[33] M. Nouh and J. R. C. Nurse, "Identifying key-players in online activist groups on the Facebook social network," in *Proc. IEEE Int. Conf. Data Mining Workshop (ICDMW)*, Nov. 2015, pp. 969–978.
[34] K. Berahmand, N. Samadi, and S. M. Sheikholeslami, "Effect of rich-club on diffusion in complex networks," *Int. J. Mod. Phys. B*, vol. 32, no. 12, May 2018, Art. no. 1850142.
[35] K. Hajarathaiah, M. K. Enduri, S. Dhuli, S. Anamalamudi, and L. R. Cenkeramaddi, "Generalization of relative change in a centrality measure to identify vital nodes in complex networks," *IEEE Access*, vol. 11, pp. 808–824, 2023.
[36] M. Alshahrani, Z. Fuxi, A. Sameh, S. Mekouar, and S. Huang, "Efficient algorithms based on centrality measures for identification of top-K influential users in social networks," *Inf. Sci.*, vol. 527, pp. 88–107, Jul. 2020.
[37] F. Grando, L. Z. Granville, and L. C. Lamb, "Machine learning in network centrality measures: Tutorial and outlook," *ACM Comput. Surv.*, vol. 51, no. 5, pp. 1–32, Sep. 2019.
[38] J. Jeyasudha and G. Usha, "An intelligent centrality measures for influential node detection in COVID-19 environment," *Wireless Pers. Commun.*, vol. 127, no. 2, pp. 1283–1309, Nov. 2022.
[39] J. Saramäki, M. Kivelä, J.-P. Onnela, K. Kaski, and J. Kertész, "Generalizations of the clustering coefficient to weighted complex networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 75, no. 2, p. 027105, Feb. 2007.
[40] W. S. Noble, "What is a support vector machine?" *Nature Biotechnol.*, vol. 24, no. 12, pp. 1565–1567, Dec. 2006.
[41] S. Suthaharan, *Machine Learning Models and Algorithms for Big Data Classification: Thinking With Examples for Effective Learning*. Cham, Switzerland: Springer, 2016, pp. 207–235.
[42] A. Navada, A. N. Ansari, S. Patil, and B. A. Sonkamble, "Overview of use of decision tree algorithms in machine learning," in *Proc. IEEE Control Syst. Graduate Res. Colloq.*, Jun. 2011, pp. 37–42.
[43] J. M. Hilbe, *Logistic Regression Models*. Boca Raton, FL, USA: CRC Press, 2009.
[44] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "KNN model-based approach in classification," in *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE*. Berlin, Germany: Springer, 2003, pp. 986–996.
[45] M. R. Segal, "Machine learning benchmarks and random forest regression," Center Bioinf. Mol. Biostatist., Univ. California, San Francisco, Tech. Rep., 2004.
[46] G. I. Webb, E. Keogh, and R. Miikkulainen, "Naïve Bayes," in *Encyclopedia of Machine Learning*, vol. 15. Berlin, Germany: Springer, 2010, pp. 713–714.
[47] J. Tang, C. Deng, and G.-B. Huang, "Extreme learning machine for multilayer perceptron," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 4, pp. 809–821, Apr. 2016.
[48] D. Bora and A. Gupta, "A comparative study between fuzzy clustering algorithm and hard clustering algorithm," *Int. J. Comput. Trends Technol.*, vol. 10, no. 2, pp. 108–113, Apr. 2014.
[49] B. Mahesh, "Machine learning algorithms—A review," *Int. J. Sci. Res.*, vol. 9, pp. 381–386, Jan. 2020.
[50] J. Wang, H. Mo, F. Wang, and F. Jin, "Exploring the network structure and nodal centrality of China's air transport network: A complex network approach," *J. Transp. Geography*, vol. 19, no. 4, pp. 712–721, Jul. 2011.
[51] L. Tahmooresnejad and C. Beaudry, "The importance of collaborative networks in Canadian scientific research," *Ind. Innov.*, vol. 25, no. 10, pp. 990–1029, Nov. 2018.
[52] J. Zhan, S. Gurung, and S. P. K. Parsa, "Identification of top-K nodes in large networks using Katz centrality," *J. Big Data*, vol. 4, no. 1, pp. 1–19, Dec. 2017.

[53] P. Jia, J. Liu, C. Huang, L. Liu, and C. Xu, "An improvement method for degree and its extending centralities in directed networks," *Phys. A, Stat. Mech. Appl.*, vol. 532, Oct. 2019, Art. no. 121891.
[54] R. A. Rossi and N. K. Ahmed, "The network data repository with interactive graph analytics and visualization," in *Proc. AAAI*, 2015, pp. 4292–4293.
[55] K. Hajarathaiah, M. K. Enduri, S. Anamalamudi, A. Abdul, and J. Chen. (2023). *Machine Learning With Centrality*. [Online]. Available: https://github.com/endurimuralikrishna/Machinelearningwithcentrality.git

KODURU HAJARATHAIAH received the B.Tech. degree in computer science and engineering from JNTUH, Hyderabad, India, in 2007, the M.Tech. degree in computer science and engineering from JNTUA, Andhra Pradesh, India, in 2010, and the Ph.D. degree in computer science and engineering from SRM University-AP, Amaravati, India, in 2023. He is currently an Assistant Professor with the Department of Computer Science and Engineering, VIT-AP University, Amaravati. His research interests include complex networks and analysis of algorithms, the information-centric networks (ICN) with IoT, and machine learning.

MURALI KRISHNA ENDURI (Member, IEEE) received the M.Sc. degree in mathematics from Acharya Nagarjuna University, Andhra Pradesh, India, in 2008, the M.Tech. degree in systems analysis and computer applications from the National Institute of Technology Karnataka, India, in 2011, and the Ph.D. degree in computer science and engineering from the Indian Institute of Technology Gandhinagar, Gujarat, India, in 2018. He is currently an Assistant Professor with the Department of Computer Science and Engineering, SRM University-AP, Amaravati, India. He has research experience/postdoctoral research experience with the Indian Institute of Technology Madras, Tamil Nadu, India, in 2018. His research interests include algorithms, complexity theory, and complex networks. Within these general areas, he has a research experience in designing algorithms for graph isomorphism problems for restricted graph classes, measuring the impact of diversity of an article, and predicting the spread of disease.

SATISH ANAMALAMUDI received the B.Eng. degree in computer science and engineering from Jawaharlal Nehru Technological University, Hyderabad, India, the M.Tech. degree in network and internet engineering from Karunya University, Coimbatore, India, and the Ph.D. degree in communication and information systems from the Dalian University of Technology, Dalian, China. He was a Research Engineer with Beijing Huawei Technologies, Beijing, China, from November 2015 to January 2017. Currently, he is an Associate Professor with the Department of Computer Science and Engineering, SRM University-AP, India. His research interests include common-control-channel design for MAC and routing protocols in cognitive radio ad-hoc networks, the MAC and routing protocol design of IoT, and 5G networks.

ASHU ABDUL (Member, IEEE) received the B.Tech. degree in computer science and engineering from Jawaharlal Nehru Technological University, Hyderabad, India, in 2009, the M.Tech. degree in computer science and engineering from the International Institute of Information Technology, Bhubaneswar, India, in 2012, and the Ph.D. degree in computer science and information engineering from Chang Gung University, Guishan, Taoyuan, Taiwan, in 2019. He is currently an Assistant Professor with the Department of Computer Science and Engineering, SRM University-AP, Amaravati, India. His main research interests include the design, analysis, and implementation of artificial intelligence algorithms and deep learning approaches.

JENHUI CHEN (Senior Member, IEEE) received the B.S. and Ph.D. degrees in computer science and information engineering (CSIE) from Tamkang University, Taipei, Taiwan, in 1998 and 2003, respectively. Since 2003, he has been with the Department of CSIE, College of Engineering, Chang Gung University, Taiwan, where he is currently a Full Professor and the Chairperson. He is also the Section Head of Technology Foresight, Artificial Intelligence (AI) Research Center, Chang Gung University. He is also a Professor with the Center for AI in medicine, Chang Gung Memorial Hospital, Taoyuan, Taiwan. His main research interests include the design, analysis, and implementation of human-like intelligence (HI), computer vision, image processing, deep learning, data science, and multimedia processing. He served as the Technical Program Committee (TPC) Member for IEEE Globecom, IEEE VTC, IEEE ICC, IEEE ICCC, IEEE ICCCN, IEEE 5G World Forum, and ACM CCIOT. He also served as a reviewer for many famous academic journals which are organized by ACM, Elsevier, IEEE, and Springer. He is currently a Senior Editor of *Cogent Engineering*.