International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 3
DOI: https://doi.org/10.17762/ijritcc.v11i3.6327
***Article Received**: 24 December 2022 **Revised**: 28 January 2023 **Accepted**: 12 February 2023*

Embedding-based Method for the Supervised Link Prediction in Social Networks

Mohamed Badiy[1], Fatima Amounas[2]
[1]RO.AL&I Group, PHD student,
Faculty of Sciences and Technics, MoulayIsmaïl University of Meknes,
Errachidia, Morocco.
e-mail: m.badiy@edu.umi.ac.ma
[2]R.O.AL&I Group, Computer Sciences Department,
Faculty of Sciences and Technics, MoulayIsmaïl University of Meknes,
Errachidia, Morocco.
e-mail: f.amounas@umi.ac.ma

Abstract— In recent years, social network analysis has received a lot of interest. Link prediction is an important area of research in this field that uses information from current networks to predict the likely links that will emerge in the future. It has attracted considerable attention from interdisciplinary research communities due to its ubiquitous applications in biological networks, computer science, transportation networks, bioinformatics, telecommunication networks, and so on. Currently, supervised machine learning is one of the critical techniques in the link prediction task. Several algorithms have been developed by many authors to predict the future link in the network, but there is still scope to improve the previous approaches. In the supervised link prediction process, feature selection is a crucial step. Most existing algorithms use one type of similarity-based feature to represent data, which is not well described due to the large scale and heterogeneity of social networks. One of the newest techniques for link prediction is embedding methods, which are used to preparing the feature vector for each the nonexisting links in the network. In this paper, we introduce a novel approach to supervised link prediction based on feature embedding methods in order to achieve better performance. Our contribution considers a set of embedding methods as the feature vector for training the machine learning classifiers. The main focus of this work is to investigate the relevance of different feature embedding methods to improve the performance of the supervised link prediction models. The experimental results on some real-world temporal networks revealed satisfactory results, which encourage us for further analysis. Moreover, the use of feature embedding methods will provide better performance in this regard.

Keywords- Social network analysis, Link prediction, Supervised learning, Feature extraction, Feature embedding methods.

I. Introduction

In the last few years, social networks have become increasingly crucial in people's lives. Social network is a popular way to model the interaction among the people in a group or community. It can be visualized as a graph, where a node corresponds to a person in that group and an edge represents some form of association between the corresponding persons. The study of social networks is the subject of a large field of research known as social network analysis (SNA). Social Network Analysis is a broad field of research which allows us to analyze and determine the relations between the nodes in the social network. Link prediction is the current trend in analyzing the social network, which exploits existing network information like the characteristics of the nodes and edges to predict the potential links that will be formed in the future [1-3]. Link prediction has many potential applications in various domains. For example, in complex network analysis, link prediction is a powerful technique to analyze complex network structure accurately [4-6]. In e-commerce, link prediction can be used for recommendations like a commodity recommendation to customers [7]. Link prediction can help in the security field by locating elusive terrorist or criminal networks [8]. In co-authorship networks, link prediction can suggest new collaborations [9]. In biological networks, link prediction can also identify the new interactions between nodes, such as protein-protein interaction networks, metabolic networks, and disease-gene networks [10, 11]. The link prediction issue has gotten greater attention from academics due to its wide range of applications in several fields. To address this issue, several link prediction approaches have been proposed in the literature. According to the taxonomy of link prediction methods presented in [12], these approaches can be classified into three categories: similarity-based methods, embedding-based methods, and learning-based methods. The network structure-based similarity method mainly contains four categories, i.e., local approaches, global approaches, quasi-local approaches, and community-based approaches. The similarity-based approach is the most frequently used approach for link prediction, which allocates similarity scores to node pairs according to the structure features of networks. Each node pair is assigned an index, defined as the similarity score between two nodes, and then it is assumed that a node pair with a higher similarity score will establish a link in the future. There are many metrics to compute the similarity between two nodes, including the common neighbor (CN), Jaccard coefficients (JC), preferential attachment (PA), resource allocation (RA), and so on [13]. These metrics used node information, local and global routes, and previous knowledge of a complex network to predict the links. Although these techniques are computationally effective, they perform less well on large real-world networks. Subsequently, the network embedding methods were proposed as variants of the similarity-based methods [14]. Sometimes the presence of embedding methods is fundamental to improving link prediction accuracy. The application of embedding methods to link prediction is a recent research trend that has attracted increasing attention from the researchers [15, 16]. The network embedding method aims to embed each node in the network into a low-dimensional feature vector while preserving the strength of the connection between nodes. This representation is used successfully in link prediction on a variety of networks [17-18], including biological networks, collaborative networks, and social networks. With the development of embedding methods research, there are many algorithms such as Deep walk based on the concept of Random walk for embedding generation [19]. Node2vec based on Deepwalk improvement [20], LINE based on simple neural network [21], and Struc2vec based on spatial similarity [22].

Recently, Machine-learning strategies are exploited in network link-prediction methods. The learning-based method develops a model based on training data and observes predicted new links. The effectiveness of machine learning techniques has been demonstrated in several practical applications, including image classification, natural language processing, and link prediction, etc. Supervised link prediction is one of the popular approaches for classifying the potential positive and negative links in the network. Although several research works have shown that this approach provides the best results, there is still a need to improve the previous approaches. Many experiments with this approach have shown promising results, but the choice of a set of features for the training of a classifier remains a major challenge. Currently, embedding methods were a useful feature

tion is embedding methods to generate the feature vector of each node of the graph

and find unknown connections.

set that contributed to higher performance [23]. This research paper suggests an effective approach to supervised link prediction based on feature embedding methods. The motivation behind the proposed approach is to improve the link prediction accuracy by using learned embedding methods as feature inputs for supervised machine learning algorithms.

Our main contributions can be summarized as follows:

- To map the problem of link prediction to the supervised machine learning domain, where the possibility of the appearance of future links is being predicted.
- To suggest a novel approach of supervised link prediction that utilizes the embedding methods as the feature vectors for training machine learning classifiers.
- To investigate the effect of network embedding methods on the performance of supervised link prediction in social networks.
- To evaluate the performance of different machine learning models by considering evaluation measures like AUC and accuracy.

The rest of this paper is organized as follows: In Section 2, we discuss the state-of-the-art literature on link prediction by focusing on network embedding and learning methods. In Section 3, we give the necessary background information for a supervised link prediction. In Section 4, we describe the proposed approach. In Section 5, we first introduce the commonly used datasets and feature vector preparation. The experimental results and the analysis part have also been discussed in this section. The last section concludes the work and gives possible future directions.

II. Related work

Over the past few years, the scientific community has intensively studied the link prediction problem. Many researchers have presented several methods to solve the link prediction problem. Some approaches in the literature focus either on similarity-based or embedding-based methods for the link prediction tasks. Initially, researchers proposed link prediction approaches based on the similarity of the nodes. For instance, Zeng et al. [24] proposed a method called "common neighbors plus preferential attachment" to calculate the likelihood of a link between two nodes based on the local knowledge of the closest neighbors. The experimental results indicate that this similarity index is highly effective and efficient. Furthermore, this method outperforms other indices, like common neighbors, resource allocation index, preferential attachment index, local path index, and Katz index. Next, Zhou et al. [25] provided a thorough algorithmic analysis of the problem of removing connections to attack similarity-based link prediction, focusing on two major groups of such methods: one that relies solely on local knowledge about the target links and the other that relies on global network knowledge. After that, Mumin et al. [26] suggested a local information-based index that uses the network topology to predict interactions between links. In addition to common neighbors, this index takes advantage of their node degrees in distributing resources. They observed that their method has robust prediction accuracy compared with other local-based metrics. Another class of link prediction approaches uses embedding-based methods, which represent the network in a low-dimensional latent space. Cao et al. [27] studied the shortcomings of network embedding algorithms by comparing them with structural similarity algorithms in short-path networks. To address this shortcoming, the authors proposed a novel method for link prediction aiming to supplement network embedding algorithms with local structural information. Experiments reveal that this method performs the best in many empirical networks, especially short-path networks. Later, Wu, C. et al. [28] proposed a link prediction algorithm based on graph embedding method for aiming at the strong randomness problem of the existing link prediction index based on random walk. The results show that the proposed method has higher accuracy than other indices. Subsequently, Tripathi, Shashi Prakash, et al. [29] proposed a novel edge embedding-based method to predict missing links in a network. The authors adopt the concepts of the skip-gram model and max aggregator for edge embedding. They evaluated and analyzed the performance of the suggested method by using four actual big networks. As a result, this method performs better and is more precise in locating the missing links than the existing techniques. Recently, machine learning has extensively contributed to the development of several link prediction approaches. Many of the researchers have focused on link prediction using supervised learning methods. For example, Pecli et al. [30] first attempt to examine and contrast the outcomes of various automatic attribute selection strategies in the link prediction problem. The authors reported the outcomes of three automatic variable selection strategies (Forward, Backward, and Evolutionary) applied to the feature-based supervised learning approach in various link prediction situations. Experiments showed promising results with three strategies. Another important approach to be mentioned is given by Kumari et al. [31]. Here, the authors developed a supervised learning-based link prediction approach that takes into account the structured-based features of social networks to predict the missing links. The proposed approach has been extensively validated by comparison with other link prediction algorithms using real-world and synthetic data sets. Another work has carried out in this field by Deepanshu et al.[32]. The authors proposed a solution for finding future links in single-layer and multiplex networks by using supervised machine learning techniques. They construct a framework for creating the training and testing data sets to evaluate the effectiveness of the machine learning classifiers. The proposed framework reached a remarkable accuracy of greater than 80% on the majority of the networks. In our previous work [33], we introduced a supervised learning approach to predict future links in Facebook Page and Dolphin social networks. To achieve this goal, we have adopted various classifiers like decision trees, logistic regression, naive Bayes, and XGBoost, with a combination of local and global similarity methods as the feature vectors. The results show that all classifiers have higher performance, greater than 80% in terms of AUC, and greater than 90% in terms of accuracy. In light of the context above, this paper presents a new approach to supervised link prediction using embedding methods as the feature vectors to achieve better performance. Although many works have presented promising results with the supervised learning approach, choosing the set of features to train the classifiers is still a major challenge. In this paper we attempt to adopt the features embedding methods as the features vectors for training the machine learning classifiers thereby improving the accuracy score.

III. PRELIMINARIES

A. *Link prediction problem in graph*

Link prediction is one of the most popular research areas in social network analysis. In the field of social networks, it is an effective way to visualize social interactions among users of social networks. The Graph is one of the most widely used data structure in the social networks field. Let’s consider an undirected graph at a particular time t where nodes represent users and links represent the relationships between pair users. The link prediction consists of predicting future connections between unconnected link pairs based on existing connections at time t+1. Figure 1 shows an example of predicting a link between users. Here, the link prediction problem aims to predict the emergence of recently formed friendships between individuals. The goal is to predict new links by considering links that already exist. Link prediction can be used to identify the existing links at time t and to predict potential relationships between users in a social network at time t+1. Solid links denote the interactions between two users at time t, while dashed links denote the interactions that will emerge at t+1. Fig. 1 depicts the notion by presenting how two users will be friends. According to the following graph, users 1 and 2 are friends at time t, and users 2 and 5 are as well. At time t+1, User5 introduced User4 to User1, and the two became friends. Similarly, users 2 and 3 would become friends at time t+1.

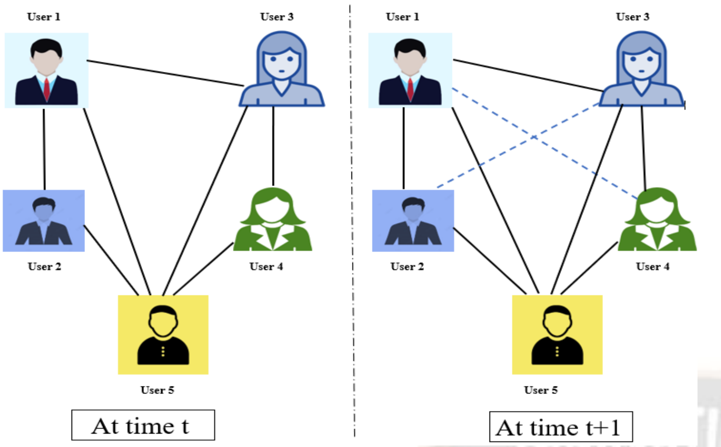


Figure 1. An example of link prediction.

B. Learning based classification

Recently, supervised machine learning has become one of the most important approaches to solving the link prediction problem. Classification is a powerful concept to deal with the link prediction problem. More precisely, the link prediction problem can be considered as a supervised binary classification problem where non-existing links are classified as positives or negatives. The main challenge of using graphs in link prediction and machine learning is finding a way to extract information about the interaction between nodes, and integrate them into a machine learning model.

Let's consider a social network represented as a graph G=(V, E), where V denotes a set of nodes and E denotes the links between them. The main goal is to find the label of the non-existing links. Let v_i and v_j be nodes in the graph G. $L^{(v_i,v_j)}$ denotes a label for the pair of nodes. If the nodes are connected, the label indicates that it is positive, otherwise, it indicates that it is negative. The label can be mathematically formulated as follows:

$$L^{(v_i,v_j)} = \begin{cases} 1, & if\ (v_i, v_j)\ \in E \\ 0, & if\,(v_i, v_j)\ \notin E \end{cases} \qquad (1)$$

In the supervised link prediction task, choosing an appropriate feature set is the most critical part of any machine learning algorithm. Extracting and defining desirable feature sets from social networks is crucial for building an effective and efficient classifier. According to Figure 2, Feature extraction techniques can be divided into two groups: (i) Features based on similarity methods, (ii) Features based on network embedding methods. The feature-based on network embedding metrics are further categorized into non-attributed and attributed network embedding metrics. Deep walk and Nodes2Vec are the most embedding methods widely used in the link prediction task.

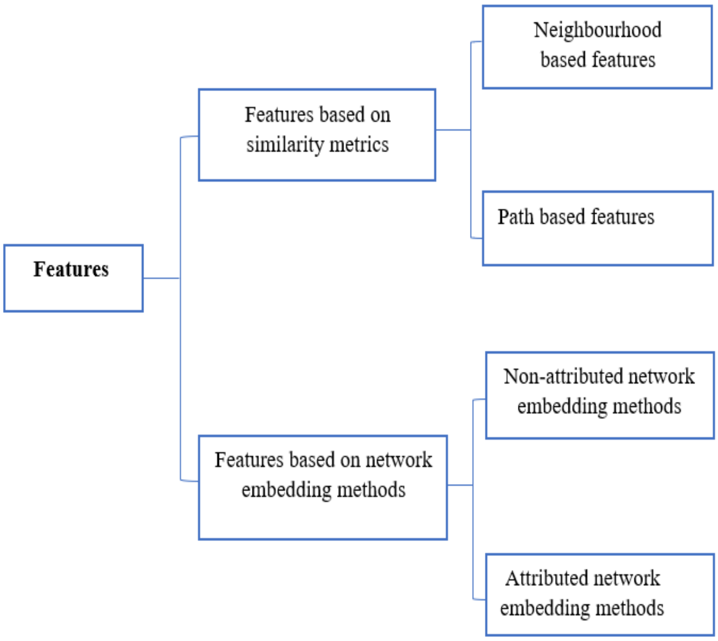

Features
Features based on similarity metrics
Neighbourhood based features
Path based features
Features based on network embedding methods
Non-attributed network embedding methods
Attributed network embedding methods

Figure 2. Features of link prediction.

IV. Proposed Methodology

With the recent growth of machine-learning strategies in various applications, supervised learning methods have been exploited in network link-prediction methods. By using an effective feature set, these approaches can achieve superior performance. This work suggests an effective approach to supervised link prediction based on feature embedding techniques. Our goal is to use embedding methods as the feature vectors for training machine learning classifiers. In this work, we employed two graph embedding algorithms: DeepWalk and node2vec. We investigated how a supervised predictor, using representations from these methods, performs on link prediction in social networks. Subsequently, we evaluate the effectiveness of supervised learning methods based on feature embedding methods for link prediction in social networks. The architecture of the proposed approach is shown in Figure 3, which consists of the three major phases of data preprocessing, feature extraction, model training, and performance evaluation.

A. Feature selection

In this section, we introduce two of the most popular embedding methods, "DeepWalk" and "Node2Vec," adopted in the proposed supervised learning model. Figure 4 depicts the node embedding process of DeepWalk and Node2Vec. The various features are described below.

- *DeepWalk* is a graph representation learning algorithm that uses random walks to generate sequences of nodes and trains a language model on these sequences to produce low-dimensional node embeddings [19]. This algorithm aims to preserve the local neighborhood information of the graph in the learned node representations, making it useful for various graph-based tasks, such as node classification and link prediction.

- *Node2Vec* is a network embedding algorithm that generates vector representations of nodes in a graph [20]. Each node is represented by a low-dimensional continuous feature vector. It aims to learn a mapping of nodes to a low-dimensional space of features using random walks through a graph starting from a target node. Unlike the Deepwalk algorithm, which uses uniform random walks, Node2Vec employs an improved biased random walk method to sample node context by considering both local and global structure information from the original graph.

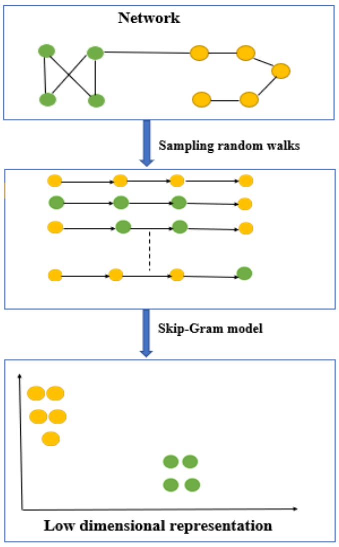


Figure 4. DeepWalk and Node2vec embedding process.

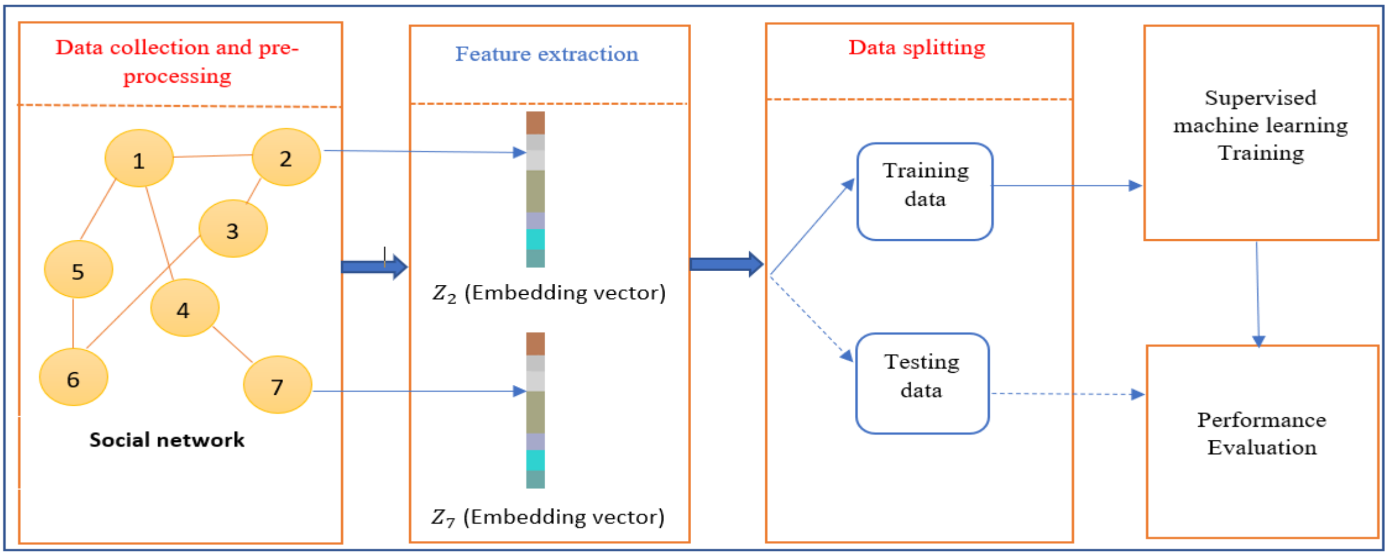


Figure 3. Block diagram of the proposed approach.

B. Supervised Machine learning algorithms adopted

The link prediction problem is one of the highlighted challenges of the social network. The usage of machine learning technology is considered a desirable technique for performing the process of supervised link prediction in order to solve this problem in the social network. The links in social networks can be predicted using a variety of machine-learning algorithms. In this work, we adopted a number of supervised machine-learning methods for the link prediction task, which are listed below:

- *Support vector machine (SVM)* is one of the most famous and widely utilized supervised learning methods for classification and regression tasks. The fundamental concept of SVM is to create a hyperplane in higher dimensional space to achieve class separation [34]. The intuition here is that the hyperplane with the most significant distance from any class's nearest training data points achieves a good separation.
- *K-nearest neighbor (K-NN)* is one of the simplest machine learning algorithms based on supervised learning techniques that can solve classification and regression problems. The basic assumption of the K-NN algorithm is that the things closest to a given data point are the most "similar" things in a data set [35]. Therefore, we can classify unforeseen points based on the values of the closest existing points. Here, the user can choose K to specify how many neighbors will be used in the algorithm.
- *Random Forest (RF)* is another supervised machine learning classifier that combines the output of multiple decision trees to reach a single result. This algorithm allows to create a more accurate and robust model [36]. It can be used to solve

regression and classification problems.

- *Adaptive Boosting (AdaBoost)* is a very popular boosting technique used as part of the ensemble method in machine learning [37]. It is the first truly effective boosting algorithm developed for the purpose of binary classification that combines several weak learners into a single strong learner. Here the weak learners are decision trees with a single split, called "decision stumps". Regression and classification issues may both be solved with AdaBoost algorithms.
- *Gradient Boosting (GB)* is another boosting technique that combines a group of relatively weak learning prediction models to create a strong predictive model [38]. These weak learning methods are usually decision trees. The basic idea behind gradient boosting is to combine multiple decision trees in order to reduce model error. Furthermore, this technique is becoming popular because of its effectiveness in classifying complex datasets.
- *Extreme Gradient Boosting (XGBoost)* is one of the highest performing algorithms used for regression and classification issues uses the gradient boosting framework to produce predictions [39]. XGBoost is just a significantly better and polished version of gradient boosting. Because it is effectively optimized, it uses fewer computational resources and produces good results quickly.

C. *Performance evaluation metrics*

In order to evaluate the performance metrics of the proposed supervised learning model, the performance metrics like accuracy and Area Under the ROC Curve (AUC) have been used. These metrics are described as follows:

- Accuracy: is a commonly used metric for evaluating the performance of a machine learning model. It measures the proportion of correctly predicted links to the total number of links in the network. It is estimated as follows:

$$Accuracy = \frac{TruePositive+TrueNegative}{P + N} \qquad (2)$$

Where:

- TruePositive are the number of links that were correctly classified as positive.
- TrueNegative are the number of links that were correctly classified as negative,
- P and N are the overall negative and positive links.

- Area Under the ROC Curve (AUC): this metric is used to measure the prediction accuracy in most link prediction algorithms. Here, two links are randomly selected, one from the set of positive pairs in the test set and another from the set of non-existent links. Then, their scores are compared. If among n independent comparisons, there are n' instances when the missing link outperforms the non-existent link and n'' instances when both of them have equal scores. The AUC is calculated using Equation 3 below:

$$AUC = \frac{n'+0,5n''}{n} \qquad (3)$$

We note that the value of AUC should be about 0.5, if all the scores are randomly generated, so the degree of AUC greater than 0.5 measures how accurate the algorithm is than the randomly selected method.

V. RESULTS AND DISCUSSION

In this section, we report on the experiments conducted to evaluate the effectiveness of our proposed approach. To evaluate the performance of our approach, we performed the experiments on four real network datasets.

A. *Datasets considered*

We perform experiments on real-world networks, and their details are mentioned in Table 1. In our case, we evaluate link prediction classifiers in relation to four social network datasets: Zakary's Karate club, Dolphin, Facebook Pages, and Twitch data set. The detailed statistics of the data sets considered are presented in Table 1.

- *Zakary's Karate Club*: is one of the most frequently used real-world network data sources. The karate club network represents social interactions among 34 individuals who were members of a karate club at a university. Wayne Zachary obtained the network data by observing social interactions among the members for a period of three years, from 1970 to 1972 [40]. Each member of the club is represented by a node, and connections among the members of the club are represented by edges.
- *Dolphin*: is an undirected and unweighted social network of bottlenose dolphins. The date of the origin of dolphins is between 1994 and 2001. In this data set, the node represents bottlenose dolphins in the bottlenose dolphin community, and an edge represents frequent associations between them. This network was made available by Lusseau et al. in 2003 [41].
- *Facebook Pages:* is an unweighted, undirected social network in which the pages are the nodes and an edge is represented by an associated link between two nodes. This network was made available by Rossi et al. in 2015 [42].
- *Twitch:* is also undirected and unweighted social network used widely by gamers to live-stream themselves while playing games. The nature of the platform is such that there are few popular gamers with many followers [43]. We chose

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 3
DOI: https://doi.org/10.17762/ijritcc.v11i3.6327
***Article Received**: 24 December 2022 **Revised**: 28 January 2023 **Accepted**: 12 February 2023*

the twitch dataset as there has not been much link prediction work done on this previously.

TABLE 1.SOCIAL NETWORK DATASETS CONSIDERED FOR THE EXPERIMENT

Network	**Nodes**	**Edges**	**Type**
Zakary's karate club	34	78	Undirected
Dolphin	62	159	Undirected
Facebook-Pages	620	2102	Undirected
Twitch EN	7126	35324	Undirected

B. *Feature vector preparation*

For nonexistence links, the feature vector is built using two network embedding methods, DeepWalk and Node2Vec, as described in Section 4. The main aim of these methods is to find out the hidden network features and encode such features as node embedding vectors in a low-dimensional space. Then, these node embedding vectors are given as input to the machine learning model. The flow diagram for node embedding vector preparation is presented in Figure 5. The steps used to create the node embedding vectors are outlined as follows:

Step 1. Convert the network that we have considered into edge-list format in preprocessing stage.

Step 2. For each node in the graph, a predefined number of random walks are generated by randomly traversing the graph. This process generates a sequence of nodes for each random walk.

Step 3. The skip-gram model is built using the series of nodes produced in step 2. In this model, each node is treated as a word in a corpus, and the sequence of nodes is treated as a sentence. The skip-gram model is trained to predict the nodes that are likely to occur in the same random walk sequence.

Step 4. To create node embeddings, the learnt skip-gram model is lastly employed. Each node in the graph is represented as a vector in a low-dimensional space, where the dimensions capture the structural properties of the node.

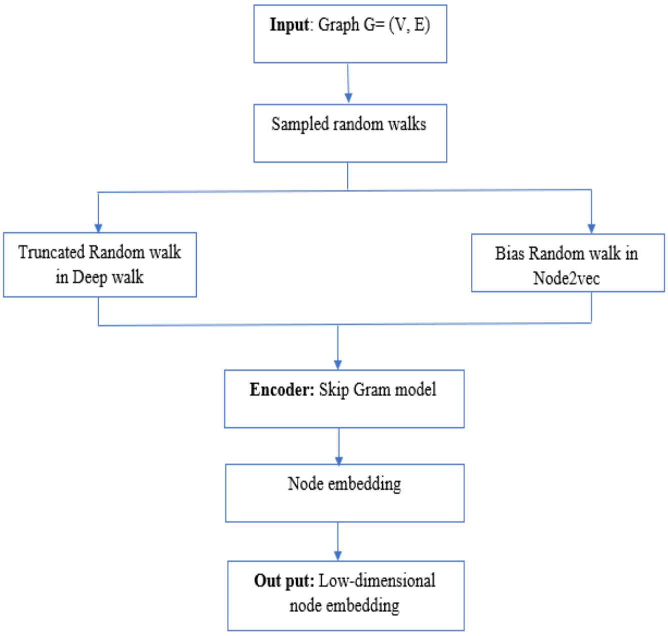

Input: Graph G= (V, E)
Sampled random walks
Truncated Random walk in Deep walk
Bias Random walk in Node2vec
Encoder: Skip Gram model
Node embedding
Out put: Low-dimensional node embedding

Figure 5. Flow diagram for node embedding features preparation

The obtained node embedding features for all nodes can be efficiently used for train the supervised machine learning algorithm. To validate the performance of our model, we should split our data into two parts: one for training the model and the other to test the model's performance. In our case, we used 70% of the data for training and the remaining 30% for testing the model, as shown in Figure 6.

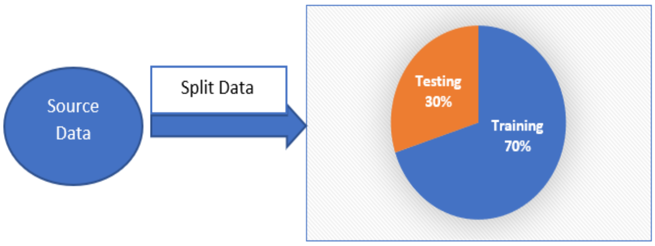


Figure 6. Training and Testing Data

C. *Experimental Results*

This section exhibits the experimental results and extensive analysis to prove the effectiveness of the proposed approach. After preparing our data and extracting our favorite features, we test it by conducting a series of experiments on several data sets of networks. Deepwalk and Node2Vec are two embedding algorithms, while PA, RA, CN and JC are the four similarity features utilized for this experiment. As described in Section 4, we use two metrics Accuracy and AUC to evaluate the performance of different classifiers utilized in this work. We also compare the accuracy scores and the AUC values obtained by supervised learning classifiers on the Zakary's Karate Club, Dolphin, Facebook Pages, and Twitch data sets. Tables 2 and 3 show the comparison of the six supervised learning classifiers, respectively.

- **Discussion**

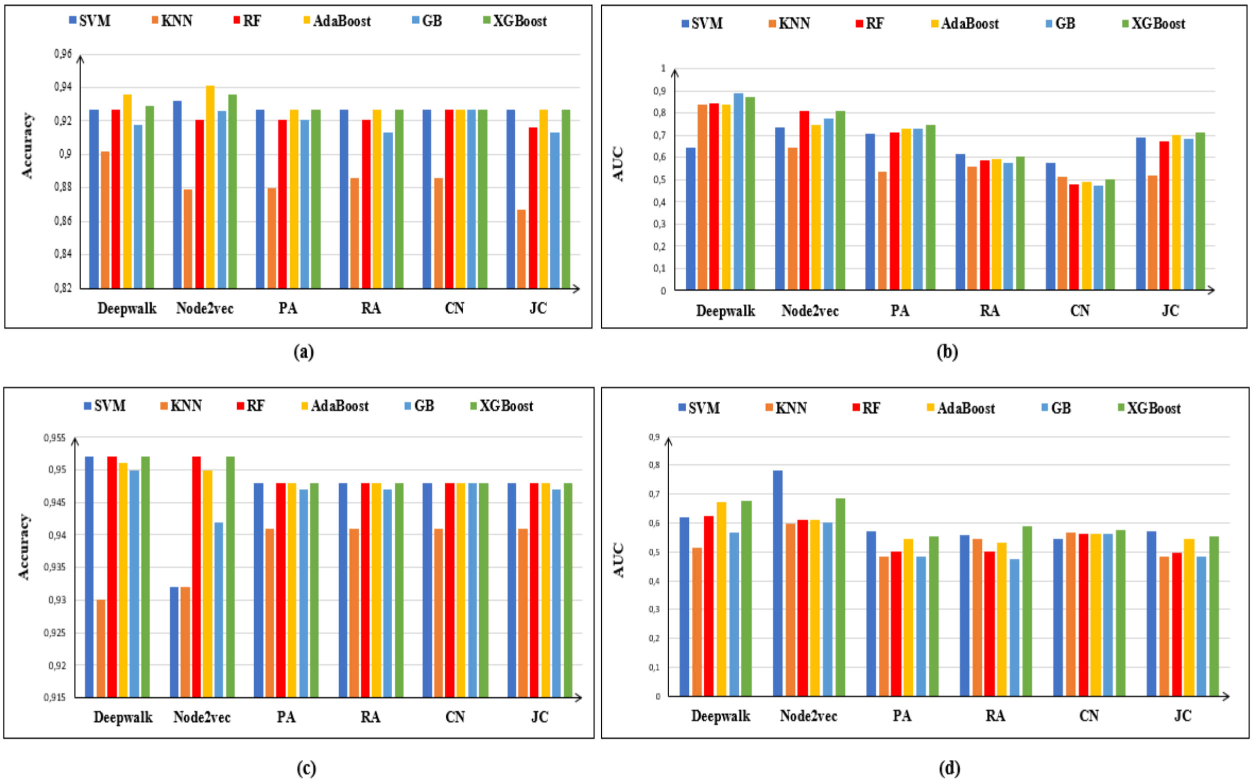
A Comparative analysis has been carried based on different metrics. Figures 7 and 8 display a visual depiction of the tables 2 and 3 discussed above. Figures 7a and 7b show the performance of several supervised machine learning models employing network embedding-based features and similarity-based features on Zakary's Karate Club dataset. According to the obtained results, when compared to other similarity-based measures, ensemble machine learning models trained with network embedding-based approaches had higher accuracy and AUC values. It can be observed that the AdaBoost model with the features DeepWalk and Node2Vec performed best, with accuracy values of 0.936 and 0.941, respectively. However, the GB model with the DeepWalk feature has a higher AUC of 0.888. In addition, it is clear that KNN is less accurate than other machine learning models in predicting links. In figures 7c and 7d, we also contrast the accuracy and AUC values of supervised machine learning models on the Dolphin dataset. The findings lead us to the following conclusions: In terms of accuracy values, models trained using DeepWalk and Node2Vec representation performed somewhat better than those trained with similarity-based metrics, while the KNN model has the lowest accuracy values. However, the best performance was seen in the SVM model with the feature of Node2Vec representations, with an AUC of 0.782. The accuracy and AUC comparison of link prediction algorithms on the Facebook pages dataset is shown in Figures 8e and 8f. It is evident from the findings that the KNN, SVM, and XGBoost algorithms trained using DeepWalk and Node2Vec produce the highest accuracy ratings. Following, all models with similarity-based features have values for accuracy and AUC that are roughly equivalent, namely 0.927 and 0.50, respectively. The comparison of link prediction models on the Twitch dataset is shown in Figures 8g and 8h. From the results, the AUC values of all models have been significantly increased. It can be observed that all models with the features DeepWalk and Node2Vec have AUC values greater than 96%, and all models with the features PA, RA, CN, and JC have AUC values greater than 70%. However, SVM outperformed with the highest AUC of 98% when tested with the Node2Vec feature. It is also seen that all models with features RA, CN, and JC have nearly identical accuracy and AUC values. Overall, the embedding feature produces better performance with various supervised learning algorithms on different network data sets. This shows the effectiveness of feature set in terms of performance.

TABLE 2. COMPARISON OF DIFFERENT MODELS IN TERMS OF ACCURACY

Dataset	**Classifier**	**Features**					
		DeepWalk	**Node2Vec**	**PA**	**RA**	**CN**	**JC**
Zakary's karate club	SVM	0.927	0.932	0.927	0.927	0.927	0.927
	KNN	0.902	0.879	0.880	0.886	0.886	0.867
	RF	0.927	0.921	0.921	0.921	0.927	0.916
	AdaBoost	0.936	0.941	0.927	0.927	0.927	0.927
	GB	0.918	0.926	0.921	0.913	0.927	0.913
	XGBoost	0.929	0.936	0.927	0.927	0.927	0.927
Dolphin	SVM	0.952	0.932	0.948	0.948	0.948	0.948
	KNN	0.930	0.932	0.941	0.941	0.941	0.941
	RF	0.952	0.952	0.948	0.948	0.948	0.948
	AdaBoost	0.951	0.950	0.948	0.948	0.948	0.948
	GB	0.950	0.942	0.947	0.947	0.948	0.947
	XGBoost	0.952	0.952	0.948	0.948	0.948	0.948
Facebook-Pages	SVM	0.942	0.944	0.927	0.927	0.927	0.927
	KNN	0.947	0.946	0.927	0.926	0.927	0.927
	RF	0.937	0.935	0.927	0.927	0.927	0.927
	AdaBoost	0.929	0.933	0.927	0.927	0.927	0.927
	GB	0.934	0.933	0.927	0.927	0.927	0.927
	XGBoost	0.942	0.940	0.927	0.927	0.927	0.927
Twitch EN	SVM	0.934	0.941	0.845	0.735	0.771	0.765
	KNN	0.921	0.925	0.896	0.770	0.771	0.770
	RF	0.931	0.931	0.897	0.771	0.771	0.771
	AdaBoost	0.910	0.919	0.897	0.771	0.771	0.771
	GB	0.924	0.927	0.897	0.771	0.771	0.771
	XGBoost	0.920	0.930	0.899	0.771	0.771	0.771

UNDER DIFFERENT DATASETS

DOI: https://doi.org/10.17762/ijritcc.v11i3.6327
***Article Received**: 24 December 2022 **Revised**: 28 January 2023 **Accepted**: 12 February 2023*



(a) (b)

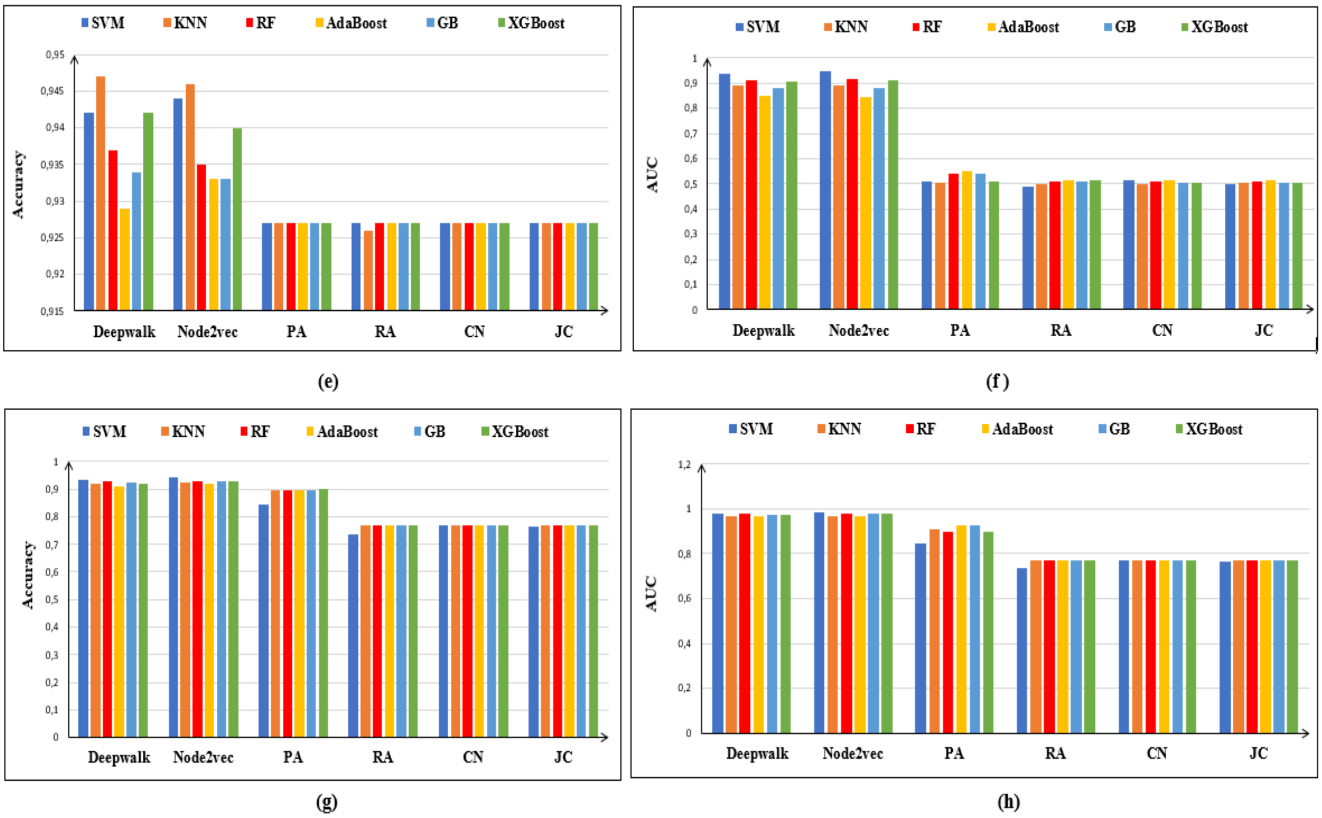
(c) (d)

Figure 7. Comparative analysis of several link prediction algorithms on Zakary's Karate Club and Dolphin data sets

TABLE 3. COMPARISON OF DIFFERENT MODELS IN TERMS OF AUC UNDER DIFFERENT DATASETS

Dataset	**Classifier**	**Features**					
		DeepWalk	**Node2Vec**	**PA**	**RA**	**CN**	**JC**
Zakary's karate club	SVM	0.646	0.734	0.708	0.615	0.574	0.692
	KNN	0.838	0.643	0.538	0.560	0.511	0.517
	RF	0.846	0.809	0.715	0.586	0.479	0.674
	AdaBoost	0.840	0.745	0.728	0.592	0.490	0.700
	GB	0.888	0.776	0.728	0.575	0.474	0.685
	XGBoost	0.872	0.812	0.749	0.602	0.50	0.711
Dolphin	SVM	0.619	0.782	0.571	0.557	0.547	0.571
	KNN	0.514	0.596	0.485	0.544	0.565	0.485
	RF	0.624	0.611	0.501	0.502	0.561	0.498
	AdaBoost	0.674	0.596	0.547	0.530	0.562	0.547
	GB	0.565	0.600	0.482	0.474	0.562	0.482
	XGBoost	0.676	0.686	0.555	0.590	0.575	0.555
Facebook-Pages	SVM	0.938	0.946	0.512	0.489	0.514	0.50
	KNN	0.889	0.892	0.504	0.498	0.498	0.503
	RF	0.912	0.919	0.541	0.509	0.507	0.507
	AdaBoost	0.850	0.843	0.549	0.513	0.517	0.517
	GB	0.880	0.881	0.538	0.507	0.504	0.504
	XGBoost	0.905	0.914	0.508	0.514	0.503	0.506
Twitch EN	SVM	0.981	0.985	0.845	0.735	0.771	0.766
	KNN	0.967	0.967	0.908	0.771	0.771	0.771
	RF	0.979	0.981	0.897	0.771	0.771	0.771
	AdaBoost	0.968	0.969	0.924	0.771	0.771	0.771
	GB	0.975	0.977	0.924	0.771	0.771	0.771
	XGBoost	0.975	0.977	0.899	0.771	0.771	0.771

DOI: https://doi.org/10.17762/ijritcc.v11i3.6327
***Article Received**: 24 December 2022 **Revised**: 28 January 2023 **Accepted**: 12 February 2023*



(e)

(f)

(g)

(h)

Figure 8. Comparative analysis of several link prediction algorithms on Facebook Pages and Twitch data sets

VI. CONCLUSION

Link prediction has been used in many fields of science, such as online social networks where links can be considered as promising friendships. The feature extraction step is one of the most crucial phases in the model for predicting the no-existing links. In this paper, we proposed an approach to supervised link prediction based on embedding methods that predicts the future link in a social network. We evaluated the effectiveness of feature embedding methods in link prediction using supervised learning. Through our experiments, we could see that the network embedding methods are a very effective feature for link prediction. From the findings, we can observed that the highest AUC value has been provided by SVM model trained with Node2vec representations on Twitch dataset. According to the obtained results, we found that machine learning models developed using network embedding methods outperformed those developed using similarity-based techniques. For future work, deep learning algorithms can be implemented in addition to ensemble machine learning methods for supervised link prediction in order to extend the analysis in this study.

References

[1] Haghani, Sogol, and Mohammad Reza Keyvanpour. "A systemic analysis of link prediction in social network.", Artificial Intelligence Review, vol. 52, 2019, pp. 1961-1995.

[2] Luo, Hongsheng, et al. "Link prediction in multiplex networks using a novel multiple-attribute decision-making approach», Knowledge-Based Systems, vol. 219, 2021, pp. 106904.

[3] Tofighy, Sajjad, Nasrollah Moghadam Charkari, and Foad Ghaderi. "Link prediction in multiplex networks using intralayer probabilistic distance and interlayer co-evolving factors." Physica A: Statistical Mechanics and its Applications, vol. 606, 2022, pp. 128043.

[4] Nasiri, Elahe, Kamal Berahmand, and Yuefeng Li. "A new link prediction in multiplex networks using topologically biased random walks." Chaos, Solitons & Fractals, vol. 151, 2021, pp.111230.

[5] Berahmand, Kamal, et al. "A preference random walk algorithm for link prediction through mutual influence nodes in complex networks." Journal of king saud university-computer and information sciences, vol. 34, 2022, 5375-5387.

[6] Xie, Feng, et al. "A link prediction approach for item recommendation with complex number." Knowledge-Based Systems, vo. 81, 2015, pp.148-158.

[7] Assouli, Nora, Khelifa Benahmed, and Brahim Gasbaoui. "How to predict crime—informatics-inspired approach from link prediction." Physica A: Statistical Mechanics and its Applications, 2021, vol. 570, pp.125795.

[8] Chuan, Pham Minh, et al. "Link prediction in co-authorship networks based on hybrid content similarity metric." Applied Intelligence, vol. 48, 2018, pp. 2470-2486.

[9] Lin, Chih-Hsu, et al. "Multimodal network diffusion predicts future disease–gene–chemical associations." Bioinformatics, vol. 35, 2019, pp.1536-1543.

[10] Nasiri, Elahe, et al. "A novel link prediction algorithm for protein-protein interaction networks by attributed graph embedding." Computers in Biology and Medicine, vol.137, 2021, pp. 104772.

[11] Daud, Nur Nasuha, et al. "Applications of link prediction in social networks: A review." Journal of Network and Computer Applications, vol. 166, 2020, pp.102716.

[12] Kumar, Ajay, et al. "Link prediction techniques, applications, and performance: A survey." Physica A: Statistical Mechanics and its Applications, vol. 553, 2020, pp. 124289.

[13] Mutlu, Ece C., et al. "Review on learning and extracting graph features for link prediction." Machine Learning and Knowledge Extraction, vol. 2, 2020, pp. 672-704.

[14] Islam, Md Kamrul, Sabeur Aridhi, and Malika Smail-Tabbone. "Appraisal study of similarity-based and embedding-based link prediction methods on graphs." Proceedings of the 10th International Conference on Data Mining & Knowledge Management Process, 2021, pp. 81-92.

[15] Cui, Peng, et al. "A survey on network embedding." IEEE transactions on knowledge and data engineering, vol.31, 2018, pp. 833-852.

[16] Cai, Hongyun, Vincent W. Zheng, and Kevin Chen-Chuan Chang. "A comprehensive survey of graph embedding: Problems, techniques, and applications." IEEE transactions on knowledge and data engineering, vol.30, 2018, pp.1616-1637.

[17] Wang, Daixin, Peng Cui, and Wenwu Zhu. "Structural deep network embedding." Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. 2016, pp. 1225-1234.

[18] Ou, Mingdong, et al. "Asymmetric transitivity preserving graph embedding." Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. 2016, pp. 1105-1114.

[19] Perozzi, Bryan, Rami Al-Rfou, and Steven Skiena. "Deepwalk: Online learning of social representations." Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. 2014, pp. 701-710.

[20] Grover, Aditya, and Jure Leskovec. "node2vec: Scalable feature learning for networks." Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. 2016, pp. 855-864.

[21] Tang, Jian, et al. "Line: Large-scale information network embedding." Proceedings of the 24th international conference on world wide web. 2015, pp. 1067-1077.

[22] Ribeiro, Leonardo FR, Pedro HP Saverese, and Daniel R. Figueiredo. "struc2vec: Learning node representations from structural identity." Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. 2017, pp. 385-394.

[23] Goyal, Palash, and Emilio Ferrara. "Graph embedding techniques, applications, and performance: A survey." Knowledge-Based Systems, vol.151, 2018, pp. 78-94.

[24] Zeng, Shan. "Link prediction based on local information considering preferential attachment." Physica A: Statistical Mechanics and its Applications, vol. 443, 2016, pp. 537-542.

[25] Zhou, Kai, et al. "Attacking similarity-based link prediction in social networks." arXiv preprint arXiv:1809.08368, 2018.

[26] Mumin, Diyawu, Lei-Lei Shi, and Lu Liu. "An efficient algorithm for link prediction based on local information: Considering the effect of node degree." Concurrency and Computation: Practice and Experience, vol. 34, 2022, pp. e6289.

[27] Cao, Ren-Meng, Si-Yuan Liu, and Xiao-Ke Xu. "Network embedding for link prediction: The pitfall and improvement." Chaos: An Interdisciplinary Journal of Nonlinear Science, vol. 29, 2019, pp. 103102.

[28] Wu, Chencheng, et al. "Link prediction based on graph embedding method in unweighted networks." 2020 39th Chinese Control Conference (CCC). IEEE, 2020, pp. 736-741.

[29] Tripathi, Shashi Prakash, Rahul Kumar Yadav, and Abhay Kumar Rai. "Network embedding based link prediction in dynamic networks." Future Generation Computer Systems, vol. 127, 2022, pp.409-420.

[30] Pecli, Antonio, Maria Claudia Cavalcanti, and Ronaldo Goldschmidt. "Automatic feature selection for supervised learning in link prediction applications: a comparative study." Knowledge and Information Systems, vol. 56, 2018, pp. 85-121.

[31] Kumari, Anisha, Behera, Ranjan Kumar, Sahoo, Kshira Sagar, et al. "Supervised link prediction using structured-based feature extraction in social network. " Concurrency and Computation: practice and Experience, vol. 34, 2020, pp. e5839.

[32] Malhotra, Deepanshu, and Rinkaj Goyal. "Supervised-learning link prediction in single layer and multiplex networks." Machine Learning with Applications, vol. 6, 2021, pp.100086.

[33] Badiy, Mohamed, Fatima Amounas, and Moha Hajar. "A Novel Hybrid Approach for Improving the Accuracy of the Supervised Link Prediction Based on Graph Structure Features in Social Networks." International Conference on Business Intelligence. Springer, Cham, 2022, pp. 231-242.

[34] Pisner, Derek A., and David M. Schnyer. "Support vector machine." Machine learning. Academic Press, 2020, pp.101-121.

[35] Taunk, Kashvi, et al. "A brief review of nearest neighbor algorithm for learning and classification." 2019 International Conference on Intelligent Computing and Control Systems (ICCS). IEEE, 2019, pp. 1255-1260.

[36] Parmar, Aakash, Rakesh Katariya, and Vatsal Patel. "A review on random forest: An ensemble classifier." International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018. Springer International Publishing, 2019, pp. 758-763.

[37] Bahad, Pritika, and Preeti Saxena. "Study of adaboost and gradient boosting algorithms for predictive analytics." International Conference on Intelligent Computing and Smart Communication 2019: Proceedings of ICSC 2019. Springer Singapore, 2020, pp. 235-244.

[38] Bentéjac, Candice, Anna Csörgő, and Gonzalo Martínez-Muñoz. "A comparative analysis of gradient boosting algorithms" Artificial Intelligence Review, vol.54, 2021, pp. 1937-1967.

[39] Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016, pp. 785-794.

[40] Zachary, Wayne W. "An information flow model for conflict and fission in small groups." Journal of anthropological research, vol. 33, 1977, pp. 452-473.

[41] Lusseau, D., Schneider, K., Boisseau, O.J. et al. The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. BehavEcol Sociobiol, vol. 54, 2003, pp. 396-405.

[42] Rossi, Ryan, and Nesreen Ahmed. "The network data repository with interactive graph analytics and visualization." Proceedings of the AAAI conference on artificial intelligence. vol. 29, 2015.

[43] Rozemberczki, Benedek, Carl Allen, and Rik Sarkar. "Multi-scale attributed node embedding." Journal of Complex Networks, vol. 9, 2021, pp. cnab014.