A Powerful Approach to improve Link Prediction Accuracy in Directed Social Networks based on Ensemble Learning Models and Advanced Feature Extraction Techniques

Sheroz Khan *, Mohamed Badiy, Fatima Amounas, Mourade Azrour, Abdullah Alnajim, Abdulatif Abdulatif

Posted Date: 30 August 2024

doi: 10.20944/preprints202408.2200.v1

Keywords: Link prediction; hyperparameter tuning; feature extraction; ensemble learning models; directed networks

Article

Powerful Approach to improve Link Prediction Accuracy in Directed Social Networks based on Ensemble Learning Models and Advanced Feature Extraction Techniques

Mohamed Badiy [1,†], Fatima Amounas [2,‡], Mourade Azrour [3,‡,*], Abdullah M. Alnajim [4,‡], Abdulatif Alabdulatif [5,‡] and Sheroz Khan [6,‡]

[1] PHD, Faculty of Sciences and Technics, Moulay Ismail University of Meknes, Errachidia, Morocco
[2] RO.AL&I Group, Computer Sciences Department, Faculty of Sciences and Technics, Moulay Ismail University of Meknes, Errachidia, Morocco
[3] IDMS team, Faculty of Sciences and Techniques, Moulay Ismail University of Meknès, Morocco
[4] Department of Information Technology, College of Computer, Qassim University, Buraydah 51452, Saudi Arabia
[5] Department of Computer Science, College of Computer, Qassim University, Buraydah 52571, Saudi Arabia
[6] Department of Electrical Engineering, College of Engineering and Information Technology, Onaizah Colleges, Onaizah 56447, Saudi Arabia
* Correspondence: mo.azrour@umi.ac.ma

Abstract: Link prediction is a significant field in network science, which focuses on predicting the probability of the existence or formation of a link between nodes in a social network based on currently observed connections. The effectiveness of traditional algorithms can vary depending on the type of network, making some methods more suitable than others for specific scenarios. Recently, several efficient link prediction algorithms have been developed, demonstrating robust results in both prediction accuracy and interpretability. However, existing research has not clearly established the relationship between network characteristics and link creation mechanisms for community influence analysis and anomaly detection. The ability to predict complex networks with diverse features still requires further investigation. In light of this, we introduce a novel framework designed to combine the best features of different link prediction algorithms when applied to the network, with the aim of achieving more reliable predictions about how networks will evolve. According to the proposed framework, we first focus on the feature extraction stage. During this phase, we systematically identify and extract a comprehensive set of features from the network before moving onto the classification phase. Here, we utilize state-of-the-art ensemble learning models to assess and classify potential links within the network. By training our machine learning models on the extracted features, we can effectively predict whether a particular link is likely to form (positive link) or unlikely to form (negative link). The ML models were trained and evaluated using two datasets: Twitch and Facebook. **Results:** Additionally, we assessed their performance on these datasets by conducting specific preprocessing and hyperparameter tuning steps. The final prediction model achieves AUC values between 94% and 99.3%. This research contributes significantly to efforts aimed at enhancing link prediction in dynamic socisal network contexts, providing valuable insights into the effectiveness of different ML algorithms in predicting future connections and enhancing our understanding of network dynamics.

Keywords: Link prediction; hyperparameter tuning; feature extraction; ensemble learning models; directed networks

1. Introduction

Networks have been providing immensely popular platforms that enable nodes to interact and communicate with each other for free or under defined conditions. Network in this context may refer to an undirected graph in which vertices are connected by edges or links. Link prediction is about finding missing links in static networks or predicting the probability of the future based on the observation of existing links in dynamic networks [1,2]. Based on empirical studies, it is possible to predict new links between vertices based on the topology of a network and the properties that

 doi:10.20944/preprints202408.2200.v1

characterize the topology of networks and the evolving dependencies between interactions over time in between two nodes in order to infer social interactions by suggesting possible friends to the users or to infer novel drugs from biological networks [3]. The task of link prediction requires to examine the proximity of different pairs of nodes and the type of interactions taking place to know how frequently any two nodes interact thus finding applications in domains of biological networks and recommender systems [4]. The growing popularity of these platforms has led to the extensive use of social network data in research across various fields, including sentiment analysis to extract opinions of people out of their writings [5], product recommendation systems based on user relationships for user by Amazon, Taobao, Jingdong and AliBaba platforms [6], user interaction studies employing statistical techniques for extracting information features from shared images and textual content [7], and social relationship analysis for spreading over the internet, metabolic networks, food webs and neural networks for various scientific and academic discipleines [8]. One crucial task in social network analysis is link prediction, which aims to identify potential or missing connections between users. The task is to predict missing links and new links in the network through existing structural information present in the network. Link prediction is a versatile technique applied across a wide array of domains to forecast potential connections within networks [9]. In social networks, it supports friend recommendations and community detection, enhancing user engagement on platforms such as Facebook and LinkedIn [10].

In the field of biomedicine [11], it predicts protein-protein interactions and genetic correlations of diseases, which help in scientific discoveries and medical research [12]. E-commerce and streaming services leverage link prediction to suggest products and content, respectively, based on user behavior [13]. Knowledge graphs use them to infer missing relationships, improve information retrieval and semantic understanding. Additionally, it plays a crucial role in fraud detection by identifying suspicious transaction patterns, and in infrastructure networks, it helps improve transportation systems and power grids [14]. Academic networks benefit from predicting future research collaborations, while telecommunications use them for network optimization [15]. Even in law enforcement, link prediction helps in uncovering hidden connections within criminal networks [16]. This wide application underscores its importance in enhancing the functionality and efficiency of various complex systems.

The entities in networks could be proteins, neurons or persons connected together edges (or links) representing associations. Link predictions are aimed at suggesting healthcare procedures for survival of patients with fatal diseases, and recommending products of interest in shopping while finding key actors in criminal investigation [17,18]. Recent studies on link prediction in social networks commonly employ two broad approaches: similarity-based methods and learning-based methods. The similarity-based method works on the assumption that nodes with higher similarity scores are more likely to be connected [19,20]. It determines the degree of similarity between nodes using a function that incorporates network data, such as topology or node attributes with relevant weighing scores. This similarity measure is then applied to estimate the likelihood and level of a link between nodes. The accuracy of the prediction heavily depends on the effective selection of network structure features. The learning-based method creates a model capable of extracting features of interest from the given network topology using computational biology, machine and data mining for drug sensitivity algorithms on profiling genomic, proteomic and epigenomic datasets. It trains this model using existing information of patients' responses to different drugs based on environmental causes, genetic factors and tumor heterogenity, and then utilizes the trained model to predict the probability of links between the pairs of nodes [21,22]. The scores of the associated links are employed to gauge the closeness of connectivity between two nodes in link prediction using scores-based heuristic methods to assess similarity by considering only the immediate neighbors shared by two nodes.

In contrast, path similarity methods leverage global structural insights of networks, encompassing paths [23,24] and ant colony optimization to predict missing links in communities, to ascertain node similarity. However, structure-based methods are solely reliant on the topology of the networks. Moreover, they may not always be reliable, different networks can exhibit varying clustering and path lengths while sharing similar degree distributions. Consequently, their performance can differ across

 doi:10.20944/preprints202408.2200.v1

varying networks, making it challenging to effectively capture the underlying topological relationships between nodes.

In recent years, a significant number of learning-based algorithms such as graph neural networks (GNN), data driven deep learning methods have evolved to aim more efficiently at improving the accuracy of link prediction in various types of networks. The learning-based methods have led to graph convolutional networks (GCN) and graph attention networks (GAN) algorithms which work on assigning different weights or importance, making them one of the most sophisticated models. for developing methods of varying applications [25,26]. These algorithms have focused on extracting essential features from networks by constructing sophisticated models. Since the extracted features form the basis for precisely predicting probable linkages, the quality and relevance of these features have a significant impact on the performance of these models. Thus, one of the most important stages in the link prediction process is the feature extraction phase. It involves identifying and selecting the most informative attributes of the network, which can include node characteristics, topological properties, and interaction patterns. By accurately capturing these features, the models can better understand the underlying structure and dynamics of the network, leading to more precise predictions of non-existing links. Thus, the success of learning-based link prediction algorithms is largely determined by the robustness and comprehensiveness of the feature extraction stage. In response to this challenge, we propose a framework that revolves around feature extraction and the application of machine learning (ML) techniques to classify potential links into two categories: "will form" (positive) or "will not form" (negative). To achieve this classification, we conducted experiments employing diverse ensemble learning models such as Random Forest (RF), Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), CatBoost, ExtraTrees, and AdaBoost.

The significant contributions of this work are:

- To introduce a novel framework for link prediction using machine learning, aiming to predict the likelihood of new connections forming in a network.
- To investigate how extracted features and ensemble learning models impact the effectiveness of link prediction process.
- To determine the optimal hyperparameter values using the GridSearchCV technique.
- To achieve the highest accuracy, we utilized machine learning classifiers with the most effective hyperparameter values determined through hyperparameter tuning.
- To evaluate the performance of various ensemble machine learning models, we considered different measures like Accuracy, AUC, Recall, Precision, and F1-score.

The rest of this paper is organized with **Section 2** presenting reviews of recent research on link prediction within social networks while **Section 3** introduces fundamental theories associated with link prediction methods. **Section 4** details the design of the proposed method and presents the experimental findings. **Section 5** concludes with a summary and discussion of the results. Finally, **Section 6** draws conclusions and presents some future works.

2. Related Work

Networks make one data type, consisting of objects of interest as nodes and with edges or links signifying some form of relations between the nodes. Such data types are used in fields ranging from those in biological sciences to analysis of terrorist networks. The links can be associated with weights providing a measure of strength for each connection, and any two nodes may be connected or not. In recent years, the scientific community has intensively studied the link prediction problem in networks, particularly those evolving with time through add and drop of nodes with time, leading to numerous algorithms based on similarity methods. However, there remains room for improving these approaches.

Currently, machine learning has significantly contributed to the development of several advanced link prediction methods. For example, the authors in [27] present an innovative approach for enhancing link prediction in social networks by leveraging an ensemble of machine learning techniques. The

 doi:10.20944/preprints202408.2200.v1

authors investigate the limitations of existing link prediction methods as the process of identifying potential connections between nodes in order to forecast the growth of patterns in a structure. The authors propose a novel ensemble framework that combines multiple machine learning algorithms to improve links predictive accuracy. Through extensive experiments and analysis, the study demonstrates the effectiveness of the proposed ensemble method in various social network scenarios. The results highlight the potential of ensemble techniques to address challenges in link prediction and contribute to more robust and accurate social network analysis. In a similar previous work [28], the authors have investigated a supervised approach to link prediction in social networks using embedding-based methods. This paper presents a novel technique that utilizes network embedding to capture the structural properties of social networks and improve the accuracy of link prediction which shows if the link exists else it helps if a link will appear between two nodes in future link prediction which can be either periodic or non-periodic in dynamic networks.

By embedding the nodes of the social network into a continuous vector space, the method allows for more effective prediction of future links based on the learned representations. The study demonstrates the effectiveness of the embedding-based approach through experiments on various social network datasets, showcasing its potential to outperform traditional link prediction methods. Another important approach to be mentioned is given by authors in [29] who propose a novel link prediction approach in complex networks by integrating recursive feature elimination (RFE) and stacking ensemble learning. This method utilizes RFE to select the most relevant features and employs a two-level stacking ensemble model combining logistic regression, gradient boosting decision tree (GBDT), and XGBoost as foundational classifiers. Their approach leverages both global and local topological information to enhance the prediction accuracy and robustness across various network datasets.

The authors in [30] propose a novel approach that combines the concept of "mean received resources" with various machine learning techniques to measure the similarity between nodes for improved accuracy of link prediction models through network structure and node characteristics. They argue that traditional link prediction methods often overlook the importance of resource-related factors that can significantly influence the formation of links in social networks. By incorporating these metrics into machine learning algorithms, the study aims to address the limitations of existing deep learning based prediction models to provide a more nuanced and effective prediction model. Another significant contribution to this field is suggested by the authors in [31] discussing network dynamic link prediction which has emerged as a powerful tool in various fields. The authors investigate link prediction techniques through supervised learning approaches, examining methods for forecasting link formation in both single-layer and multiplex networks. While single-layer networks involve a single type of connection, multiplex networks feature multiple types or layers of connections. This study emphasizes how applying supervised learning models to these diverse network structures can enhance the accuracy of link prediction. Moreover, Ghorbanzadeh Hossien, et al. [32] introduced an innovative method that combines multiple techniques to improve the accuracy of predicting future links in directed graphs, where edges have a specific direction. Their approach integrates various prediction strategies to capitalize on different aspects of the graph's structure and dynamics. By merging the strengths of these individual techniques, this method aims to enhance overall prediction performance, as demonstrated through experiments on directed graph datasets.

The authors in [33] have investigated a supervised link prediction method that uses structure-based feature extraction in social networks. The authors have proposed a method to extract features derived from the network's structure for improving the prediction of future links. By emphasizing these structural features, their method seeks to enhance the accuracy of link prediction in social networks. The effectiveness of the approach is demonstrated through experiments conducted on various social network datasets.

Further, the authors in [34] have explored link prediction in multiplex networks by employing inherent features of recursive feature elimination of random forest to select representative and relevant

 doi:10.20944/preprints202408.2200.v1

structural features of the networks and using stacking method to enhance prediction results of the model. The authors developed a method specifically designed to predict links in networks with multiple layers or connection types by employing logistic regression (LR), gradient boosting decision tree (GBDT), and XGBoost as the base models and using XGBoost as the top-level model. Their approach leverages supervised learning to effectively navigate the complexities of multiplex networks and boost the accuracy of link predictions. The study validates the effectiveness of this strategy through extensive experimentation on multiplex network datasets. Different performance metrics are used to evaluate the methods discussed in the literature review. Table 1 provides a summary of the best classifier for each approach based on the selected metrics.

Table 1. Performance of Recent Research Cited in the Literature Review on Link Prediction.

Reference	Classifier	Performance achieved		
		Metrics	Value	Dataset
[27]	Ensemble Method	Accuracy	94.23%	Facebook
[28]	SVM	AUC	98.5%	Twitch EN
[29]	RF-RFE-SELLP	AUC	99.49%	NetScience
[30]	Improved RA	AUC	95.5%	USAir
[31]	RF	Precision	95.72%	Caida
[32]	SCN-HA	AUC	90%	Wiki-Vote
[33]	SVM	AUC	93.6%	Synthetic network 3
[34]	AdaBoost	Recall	86.98%	Vicker

3. Proposed Methodology

The proposed model employs a link prediction classification approach utilizing feature extraction and ensemble learning models. Our goal is to determine the most significant representative features for training these models. The research workflow is as shown in Figure 1 that includes several steps in the experimental process. First, step is related to gathering and preparing the dataset followed by features selection. Afterward, the seven classifier methods are evaluated with the selected features to identify the best classification approach for link prediction.

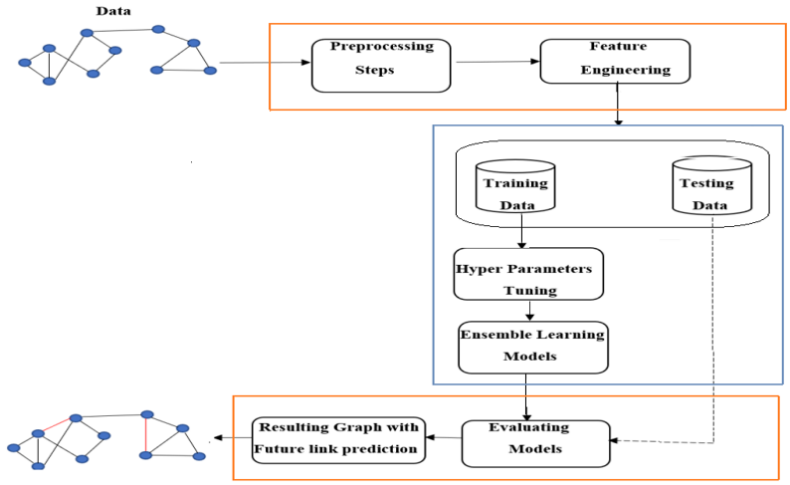

Figure 1. Flowchart of the proposed approach.

3.1. Dataset

The datasets from two social networks of Twitch and Facebook are used to assess the effectiveness of our proposed method. The choice of these platforms is driven by their distinct user engagement patterns and diverse content, providing a comprehensive framework for our analysis. Through the use of these datasets, the robustness and applicability of our method across different social media environments is demonstrated. Table 2 provides a comprehensive overview of the dataset statistics.

 doi:10.20944/preprints202408.2200.v1

1 **Twitch** is a social network widely used by gamers to live-stream their games. It allows users to watch and interact with gamers in real time, fostering a vibrant community of viewers and streamers. The platform features a small number of popular gamers who have a large number of followers, creating a highly skewed follower distribution. The **Twitch** dataset is chosen for this study because it presents a unique and under-explored opportunity for link prediction research, given the limited amount of previous works in this area. The distinct characteristics of the dataset, including the engagement patterns and follower dynamics, make it a very interesting subject for analyzing and predicting new connections.
2 **Facebook** is a widely used social network that connects people from all over the world. Users create profiles, share content, and interact with friends, family, and various communities through posts, comments, likes and messages. The platform is a complex network structure with a large number of connections and interactions, making it a rich source of data for social network analysis. Given its extensive user base and the diversity of interactions, **Facebook** network provides numerous opportunities for studying link predictions. Understanding how connections form, evolve, and influence user behavior can provide valuable insights into social dynamics and the spread of information. The Facebook dataset, with its complex network of relationships, serves as a fertile ground for developing link prediction research, particularly in studying how online interactions translate into network growth and the formation of new connections.

Table 2. Data set considered for the experiment.

Dataset	Number of nodes	Number of edges	Source
Twitch	7126	35324	[35]
Facebook	4039	88234	[36]

3.2. Feature Engineering

Defining to select a set of features in the task of link prediction for constructing the classification model is crucial for any classifier. In this study, the authors have used the edges from the datasets under consideration to extract a range of 14 features, as detailed in Figure 2 while Figure 3 offers detailed visual representations of the correlations among 14 features in the context of Twitch and Facebook datasets.

1 **Shortest Path:** In graph theory, the shortest path between two vertices (or nodes) is the route that minimizes the total weight of the edges along the path. Various algorithms can be used to determine the shortest path, depending on the characteristics of the graph and the specific needs of the problem.
2 **Page Rank_Src (Source):** It represents how much importance or "rank" the source page contributes to the linked or destination page. The higher the PageRank of the source page, the more influence it has on the rank of the destination page.
3 **Page Rank_Dst (Destination):** It refers to the PageRank value of the destination page that receives hyperlinks from other pages. It is a key component in determining the importance and visibility of a webpage in search engine results. The Page Rank_Dst value accumulates contributions from all the source pages Page Rank_Src (Destination): that link to it.
4 **Jaccard_for_followees:** Jaccard Similarity is a metric used to measure the similarity between two sets. It is defined as the size of the intersection divided by the size of the union of the two sets. When applied to followees (people or entities that users follow on a social media platform), it helps in understanding how similar the followees sets of two users are.
5 **Cosine_for_followees:** Cosine similarity is another metric used to measure the similarity between two sets or vectors. Unlike Jaccard similarity, which considers the intersection and union of sets, cosine similarity considers the angle between two vectors in a multi-dimensional space. When applied to followees, it helps in understanding how similar the followee sets of two users are by representing by representing the followees of each user as a vector.

 doi:10.20944/preprints202408.2200.v1

6 **Preferential_attachment_for_followers:** Preferential attachment is a concept from network theory that describes how the probability of a new node connecting to an existing node in a network depends on the degree (number of connections) of the existing node. In the context of social networks and followers, preferential attachment helps to explain how users with more followers are more likely to gain additional followers.

7 **katz_centrality:** is a measure used in network theory to assess the relative importance or centrality of nodes within a network. Unlike simpler centrality measures like degree centrality, which counts the number of direct connections, Katz centrality takes into account indirect connections and the quality of those connections.

8 **Adar_index:** is a measure of similarity between nodes in a network, which is often used in social network analysis and recommendation systems. It quantifies how similar two nodes are based on their shared connections and the uniqueness of those connections. Nodes that share connections through rare nodes are considered more similar than those sharing connections through common nodes.

9 **Followers_Src:** is a metric that refers to the followers of a source entity or user in the context of social media or network analysis. It denotes the set of users who follow a particular source entity or user.

10 **Followees_Src:** refers to the users or entities that a particular source entity follows in the context of social-network platforms. It represents the set of accounts, profiles, or entities that the source entity has chosen to follow.

11 **Followers_Dst:** refers to the followers of a destination entity or user in the context of social media or network analysis. It denotes the set of users who follow a particular destination entity or user.

12 **Followees_Dst:** refers to the entities or users that a destination entity or user follows in the context of social media or network analysis. It denotes the set of users or entities that are being followed by a particular destination entity or user.

13 **Int_Followers:** refer to users who are part of a specific network or community within the platform. This could include followers who are members of the same organization, group, or closed community, where certain updates or posts are limited to "internal followers" only.

14 **Int_Followees:** Internal Followees are individuals, departments, or specific entities within an organization or network whom another entity such as a department, team, or user follows or subscribes to for updates, communications, or information within the network or organisation platform.

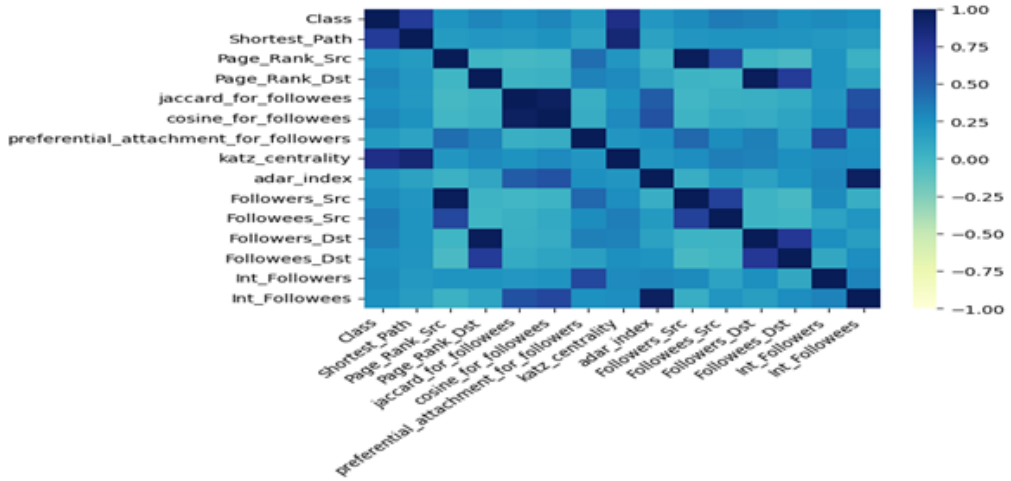

Figure 2. Correlation Among Different Variables on Twitch.

 doi:10.20944/preprints202408.2200.v1

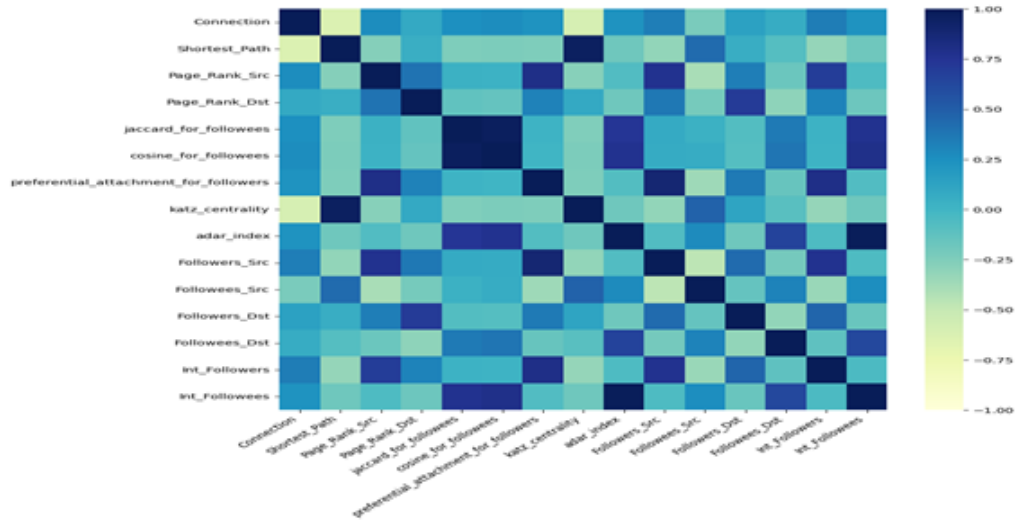

Figure 3. Correlation Among Different Variables on Facebook.

3.3. Ensemble Learning Algorithms Construction

Machine learning is a subset of artificial intelligence and computer science that focuses on training algorithms to imitate the way humans learn to identify patterns. By utilizing statistical techniques, these algorithms learn from past data and use that knowledge to make predictions or decisions using data or to generate new data [37]. One significant approach within machine learning is ensemble learning from information embedded in the networks, which integrates several models to enhance overall performance and robustness [38]. The information can be either structured which represents topological structure of the network or content based which indicate features associated with the entities and their relationships In this paper has been employed six ensemble learning models: Random Forest (RF), Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), CatBoost, ExtraTrees, and AdaBoost. We start by training the model with all available features and then assess its performance using metrics such as accuracy, AUC, precision, recall and F1-score.

Overfitting is an undesirable phenomenon that occurs in machine learning models which give accurate predictions for data they are trained with, but not for all types of new data. To prevent overfitting and evaluate the effectiveness and reliability of the various ensemble learning models [39], we utilized a 10-fold cross-validation technique along with a hyperparameter tuning process. This approach helped us identify the optimal hyperparameters, which were then applied during the training phase of each model.

3.4. Performance Metrics

Key performance indicators, such as Accuracy, F1-score, Precision, Recall, and the Area Under the Receiver Operating Characteristic Curve (AUC/ROC), are used to evaluate the suitability of each approach. In this paper the effectiveness of the method has been assessed using the metrics described in details as follows:

(a) Accuracy: is the proportion of correctly predicted links, including both existing and non-existing ones, relative to the total number of predictions. For a classification problem, accuracy can be expressed as:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

Where:

- TP = True Positives
- TN= True Negatives

 doi:10.20944/preprints202408.2200.v1

- FP = False Positives
- FN = False Negatives

(b) Precision

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

(c) Recall

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

(d) F1-Score: is the combined average of precision and recall, offering a single measure that balances both factors. The F1-Score is calculated using Equation 4 below:

$$F1\text{-}score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

(e) AUC (area under the ROC curve): refers to a performance metric used to evaluate the effectiveness of a classification model. To calculate AUC, perform the following steps: randomly choose one existing edge and one non-existent edge from the test set, and compare their scores. If the score for the existing edge is higher, increment t_1 by 1. If the scores are equal, increment t_2 by 1. Finally, the AUC can be computed as:

$$AUC = \frac{t_1 + 0.5 \times t_2}{t} \tag{5}$$

Where t represents the total number of comparisons.

4. Experimental Results and Discussion

The experimental results and comprehensive analysis described in this section is aimed at validating the effectiveness of the proposed approach. The experimental setup is outlined, detailing the procedures, and the tools used. This is followed by examination of feature importance models, which provide insights into features most significantly influencing model predictions. Analyzing these features allows us to better understand their role in model performance. Finally, an extensive discussion of the results is provided.

4.1. Experimental Setup

In this section are explained how the experiments are conducted. This study has involved the development of a machine learning classifier and the incorporation of feature selection methods using Python scripts. All experiments have been performed on a computer with 16 GB of RAM and an Intel Core i5 CPU running Windows 11. The algorithms have been tested in the Google Colab environment. The hyperparameter tuning step is performed to enhance the accuracy of the approach, selecting the best parameters for each algorithm. GridSearchCV method has been used to optimize the hyperparameters of each model. This method automates the process of selecting the best combination of parameters for a given algorithm by exhaustively searching through a specified parameter grid. Table 3 displays the best hyperparameter combinations for the algorithms used with the Twitch and Facebook datasets. The steps involved in GridSearchCV are outlined below:

1. **Define the Model:** Select the machine learning algorithm to optimize.
2. **Create the Parameter Grid:** Specify the parameters and their ranges to test, typically using a dictionary where the keys are parametric names and the values are lists of possible values.
3. **Configure GridSearchCV:** Initialize the GridSearchCV object with the model, parameter grid, and options like the number of cross-validation folds.
4. **Fit the Model:** Train the model on the training data using the specified parameter grid to cross-validate.

 doi:10.20944/preprints202408.2200.v1

5. **Evaluate the Results:** Analyze the results to identify the best parameters and to evaluate the model's performance with those parameters

Table 3. Best values for hyperparameter tuning.

	Algorithms	Best hyperparameter values
Twitch	RF	'max_depth' = 10, 'min_samples_split' = 10, 'n_estimators' = 300
	XGBoost	'colsample_bytree' = 0.8, 'learning_rate' = 0.01, 'max_depth' = 6, 'n_estimators' = 200, 'subsample' = 1.0
	LightGBM	'learning_rate' = 0.01, 'max_depth' = 10, 'n_estimators' = 100, 'num_leaves' = 31
	CatBoost	'depth' = 6, 'iterations' = 100, 'I2_leaf_reg' = 7, 'learning_rate' = 0.1
	ExtraTrees	'max_depth' = 30, 'min_samples_leaf' = 1, 'min_samples_split' = 10, 'n_estimators' = 300
	AdaBoost	'base_estimator__max_depth' = 4, 'learning_rate' = 0.1, 'n_estimators' = 50
Facebook	RF	'max_depth' = None, 'min_samples_split' = 2, 'n_estimators' = 300
	XGBoost	'colsample_bytree' = 0.8, 'learning_rate' = 0.2, 'max_depth' = 9, 'n_estimators' = 300, 'subsample' = 0.8
	LightGBM	'learning_rate' = 0.1, 'max_depth' = 10, 'n_estimators' = 300, 'num_leaves' = 100
	CatBoost	'depth' = 8, 'iterations' = 300, 'I2_leaf_reg' = 7, 'learning_rate' = 0.2
	ExtraTrees	'max_depth' = None, 'min_samples_leaf' = 1, 'min_samples_split' = 5, 'n_estimators' = 300
	AdaBoost	'base_estimator__max_depth' = 3, 'learning_rate' = 1, 'n_estimators' = 200

4.2. Feature Importance Models

"Feature importance" encompasses a range of techniques designed to assign a significance score to each input parameter based on its ability to predict a target variable [40]. These scores are integral to a predictive modeling project for several reasons. Firstly, they provide valuable insights into the dataset, revealing which features have the greatest impact on the prediction results. This understanding can inform data preprocessing steps, such as cleaning and transforming data, to improve model performance. Secondly, feature importance scores shed light on the inner workings of the model itself. By identifying the features the model relies on most, practitioners can gain a deeper understanding of the model's decision-making process. This can aid in the interpretation of models, making it easier to explain and justify predictions to stakeholders. The significance of the 14 features across the six models employed in our approach has been described in this section. Figure 4 illustrates the significance of various features in the Twitch dataset.

Specifically, Figures 4(a), Figure 4(b), and Figure 4(f) show that Followers_Dst and Followees_Src are highly significant in the RF, XGBoost, and AdaBoost models. These features consistently rank at the top across these models, underscoring their substantial impact on model performance. On the contrary, the LightGBM and CatBoost models emphasize the importance of Followees_Src and Page Rank_Dst, as depicted in Figure 4(c) and Figure 4(e). These features are critical to the prediction accuracy of LightGBM and CatBoost. In contrast, the Extra Trees model, as shown in Figure 4(d), identifies katz_centrality and Shortest Path as the most influential features, highlighting their unique contribution to the model's performance.

 doi:10.20944/preprints202408.2200.v1

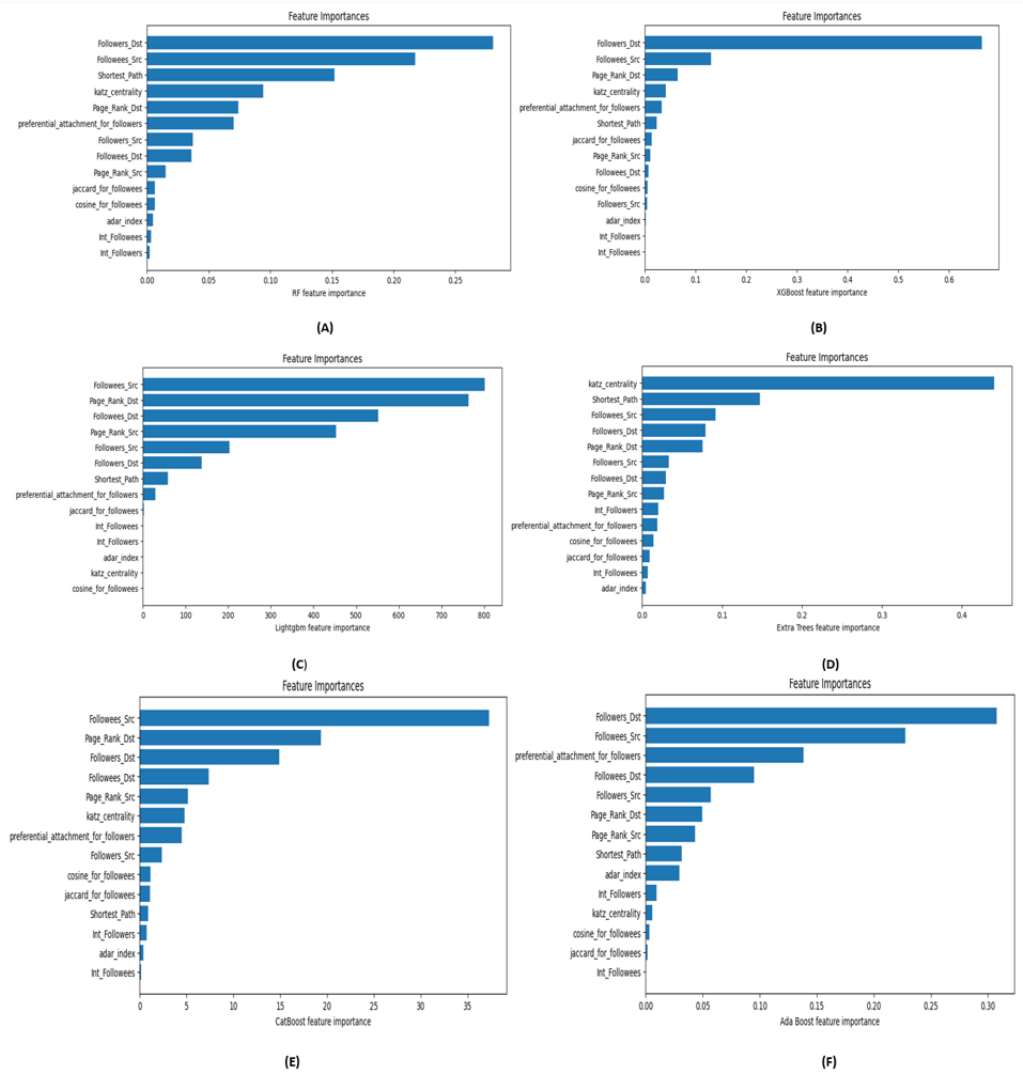

Figure 4. The feature importance plots of the six machine learning models for the Twitch dataset.

Similarly, Figure 5 demonstrates the importance of various features in the Facebook dataset. Figures 5(a), Figure 5(b), and Figure 5(d) reveal that Shortest Path and katz_centrality are highly significant in the RF, XGBoost, and Extra Trees models. These features consistently rank at the top, indicating their substantial impact on these models. On the other hand, the LightGBM and AdaBoost models prioritize Page Rank_Dst and Page Rank_Src, as shown in Figures 5(c) and Figure 5(f) highlighting their critical role in these models' prediction accuracy. For the CatBoost model, Figure 5(e) showcases katz_centrality and Followees_Src as the most influential features, suggesting their unique contribution to the model's performance compared to the others.

 doi:10.20944/preprints202408.2200.v1

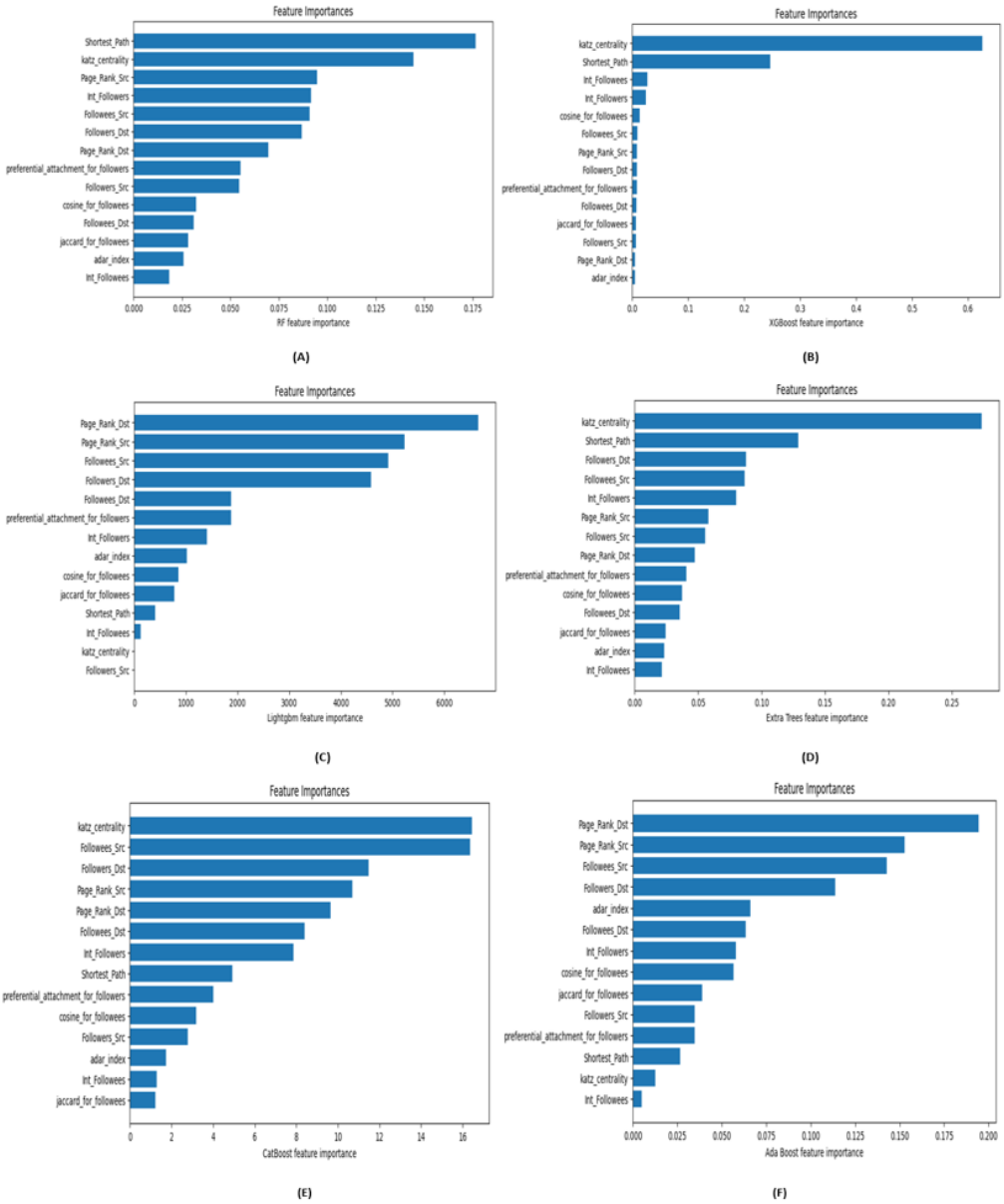

Figure 5. The feature importance plots of the six machine learning models for the Facebook dataset.

4.3. Results

The performance metrics discussed in subsection 3.4 have been used here to present the results of the classification models. Also, the developed models are trained on the training dataset to validate their performance on new data to check if they suffer from outfitting.

Figure 6 illustrates the precision of the classifiers before and after hyperparameters tuning on the Twitch and Facebook datasets. The experimental results clearly show that hyperparameter tuning enhanced the precision of most classifiers. These results are achieved by adjusting the hyperparameters to optimize precision. Using 10-fold cross-validation, the precision of different hyperparameter combinations is evaluated. These results demonstrate that most classifiers, including RF, XGBoost, CatBoost, ExtraTrees, and AdaBoost, have improved the precision with hyperparameter tuning. However, the precision of LightGBM has remained unchanged in the Twitch dataset.

 doi:10.20944/preprints202408.2200.v1

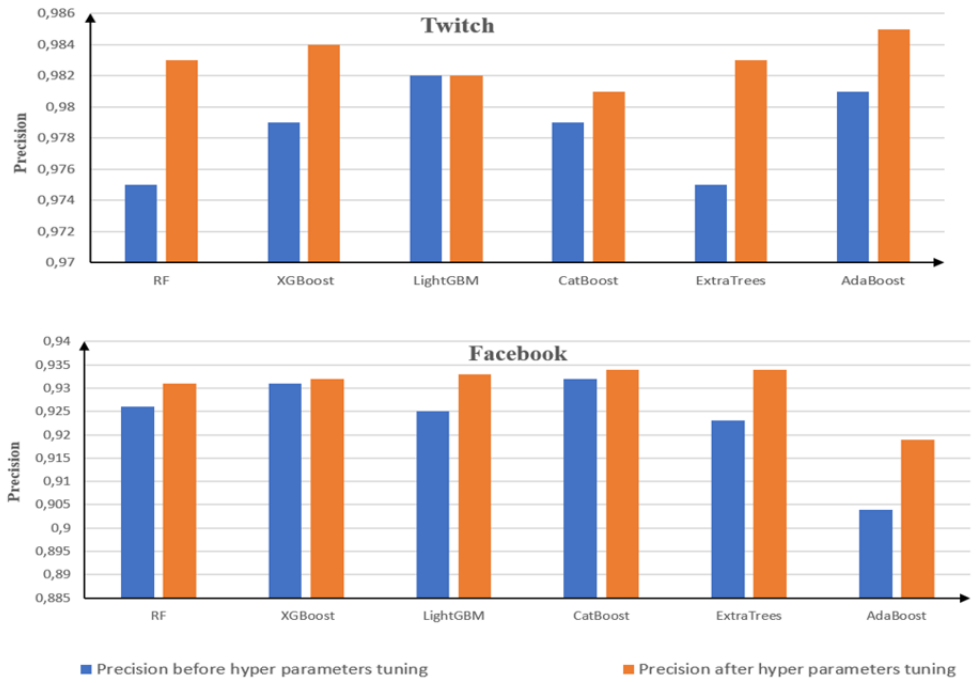

Figure 6. Precision of classifiers before and after hyper parameter tuning.

Table 4 provides a comparative analysis of the performance of different algorithms across multiple metrics, including Accuracy, AUC, Precision, Recall, and F1-Score. This evaluation reveals subtle variations in performance between classifiers when applied to the Twitch and Facebook datasets.

For the Twitch dataset, all classifiers have performed exceptionally well, with XGBoost and CatBoost slightly outperforming the others in accuracy (0.968) and AUC (0.993). Random Forest, LightGBM, and AdaBoost also demonstrated strong results with accuracy scores of 0.967 and comparable AUC values, indicating high reliability in class discrimination. It is worth noting, the precision scores for all classifiers were very high, ranging from 0.981 to 0.985, suggesting a low rate of false positives. The F1-scores were consistently around 0.968, reflecting balanced precision and recall. Although ExtraTrees still performed well, it had a marginally lower AUC of 0.981, indicating a slightly lesser but still robust distinction capability.

In contrast, the Facebook dataset presented a greater challenge, resulting in lower overall accuracy scores across all classifiers. XGBoost again has delivered with an accuracy of 0.921 and a highest AUC of 0.976, indicating superior performance in class separation. CatBoost and LightGBM followed closely, maintaining high AUC values of 0.974 and 0.972, respectively, with balanced precision and recall, resulting in strong F1-scores (around 0.930). Random Forest, while having a high AUC of 0.972, showed a slightly lower accuracy of 0.911 and an F1-score of 0.923, suggesting it may have had more difficulty with the Facebook dataset than with the Twitch dataset. ExtraTrees and AdaBoost had the lowest accuracy scores of 0.906 and 0.902, respectively, with AdaBoost showing a notably lower AUC of 0.942. Despite this, the precision scores remained relatively high, but the recall was slightly lower, reflecting a higher rate of missed positive instances. Overall, while all classifiers exhibited efficient performance, XGBoost and CatBoost consistently achieved the best results across both datasets. However, the Facebook dataset proved more challenging to accurately classify compared to the Twitch dataset. The visual representation of classifier performance based on the Area Under Curve (AUC) is as shown in Figure 7 and Figure 8.

 doi:10.20944/preprints202408.2200.v1

Table 4. Resultant analysis for various classifiers.

	Classifier	Accuracy	AUC	Recall	Precision	F1-score
Twitch	RF	0.967	0.992	0.955	0.983	0.968
	XGBoost	0.968	0.993	0.952	0.984	0.968
	LightGBM	0.967	0.993	0.954	0.982	0.968
	CatBoost	0.968	0.993	0.955	0.981	0.968
	ExtraTrees	0.966	0.981	0.955	0.983	0.968
	AdaBoot	0.967	0.993	0.950	0.985	0.967
Facebook	RF	0.911	0.972	0.914	0.931	0.923
	XGBoost	0.921	0.976	0.931	0.932	0.931
	LightGBM	0.920	0.972	0.928	0.933	0.931
	CatBoost	0.919	0.974	0.926	0.934	0.930
	ExtraTrees	0.906	0.967	0.910	0.934	0.922
	AdaBoot	0.902	0.942	0.908	0.919	0.913

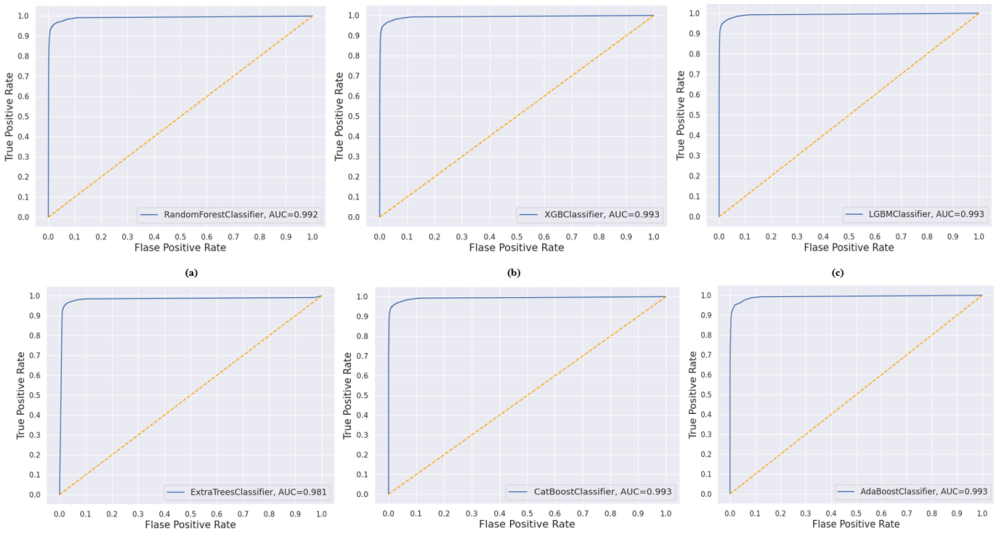

Figure 7. AUC curve for different classifiers in Twitch dataset.

 doi:10.20944/preprints202408.2200.v1

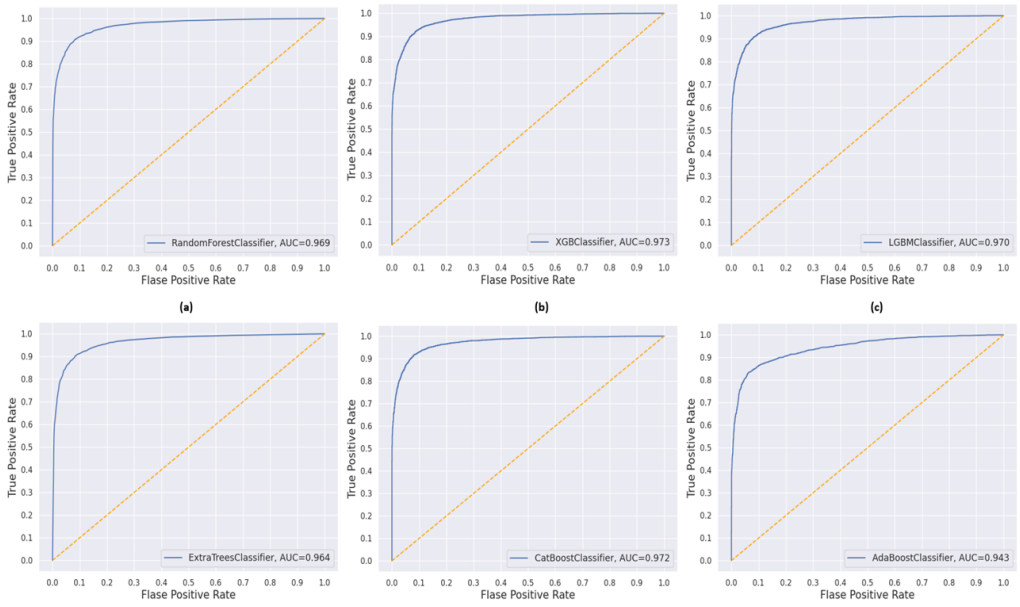

Figure 8. AUC curve for different classifiers in Facebook dataset.

4.4. Discussion

The results of analysis of the Twitch and Facebook datasets confirm the profound impact of ensemble learning models and feature selection on model performance. Ensemble methods, including XGBoost, CatBoost, and LightGBM, consistently outperformed other classifiers across both datasets, demonstrating their ability to improve accuracy, AUC, precision, and recall. These models leverage the strengths of multiple underlying learners to create a more powerful predictive system, effectively capturing complex data patterns and mitigating issues such as overfitting. For instance, both XGBoost and CatBoost have achieved the highest accuracy and AUC scores, indicating their exceptional capability in distinguishing between classes and handling diverse data complexities.

Feature selection further amplifies the effectiveness of these models by selecting and retaining only the most relevant predictors. This process eliminates noise and redundant features, enhancing model accuracy and interpretability while reducing computational load. In the context of the Facebook dataset, which posed a greater challenge, the high AUC scores and balanced metrics achieved by the ensemble methods suggest that thoughtful feature selection played a critical role in managing increased complexity and noise.

Overall, the synergy between ensemble learning and effective feature selection leads to models that are not only highly accurate, but also efficient and interpretable. This balanced approach is essential to address complex real-world problems, ensuring that models are both robust and practical for deployment.

5. Conclusions

In this paper, the authors have presented an innovative method to predict links in Twitch and Facebook networks using machine learning algorithms and engineered features. The main idea in this research is to evaluate the effectiveness of different ensemble learning algorithms to determine the most suitable approach for link prediction in these social networks. By employing Grid Search for hyperparameter optimization, the classifiers' performance is enhanced. The findings show that the XGBoost and CatBoost algorithms have achieved the highest accuracy and AUC scores, making them the most effective models for this task. These results highlight the potential of utilizing ensemble learning models combined with advanced feature extraction techniques to significantly enhance the accuracy of link prediction in directed social networks. For future research, we recommend exploring the use of Genetic Algorithm-based feature selection to further optimize the feature set

 doi:10.20944/preprints202408.2200.v1

and discover new informative features that could enhance the predictive power of the models. This evolutionary approach could provide a more automated and efficient way to optimize feature subsets, leading to even greater improvements in link prediction performance. The paper has introduced a novel framework for enhancing link prediction accuracy in directed social networks by leveraging advanced feature extraction techniques and ensemble learning models. The significant contributions of this work include: (1) the development of an innovative approach to predict the likelihood of new connections forming in a network using machine learning; (2) an in-depth investigation into the impact of feature extraction and ensemble models on the link prediction process; (3) the identification of optimal hyperparameters through the GridSearchCV technique to maximize model performance; (4) the application of machine learning classifiers with finely-tuned hyperparameters to achieve superior accuracy; and (5) a comprehensive evaluation of various ensemble models using metrics such as Accuracy, AUC, Recall, Precision, and F1-score, demonstrating the efficacy of the proposed approach in improving prediction accuracy.

Acknowledgments: The author acknowledge the support of Nawaf Waqas, PhD Scholar, Universiti Kuala Lumpur (UniKL), Malaysia, for technical support of functional plotting and to template the paper in Latex.

Conflicts of Interest: There is no conflict of interest among the authors.

References

1. Kumar, A., Singh, S. S., Singh, K., and Biswas, B. "Link prediction techniques, applications, and performance: A survey." *Physica A: Statistical Mechanics and its Applications* 553 (2020): 124289.
2. Chen, J., and Ying, R. "Tempme: Towards the explainability of temporal graph neural networks via motif discovery." *Advances in Neural Information Processing Systems* 36 (2024).
3. You, Y., Lai, X., Pan, Y., Zheng, H., Vera, J., Liu, S., Deng, S., and Zhang, L. "Artificial intelligence in cancer target identification and drug discovery." *Signal Transduction and Targeted Therapy* 7, no. 1 (2022): 156.
4. Li, J., Shomer, H., Mao, H., Zeng, S., Ma, Y., Shah, N., Tang, J., and Yin, D. "Evaluating graph neural networks for link prediction: Current pitfalls and new benchmarking." *Advances in Neural Information Processing Systems* 36 (2024).
5. Tanantong, T., Sanglerdsinlapachai, N., and Donkhampai, U. "Sentiment classification on Thai social media using a domain-specific trained lexicon." In *2020 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, IEEE, 2020.
6. Long, J., et al. "A recommendation model based on multi-emotion similarity in the social networks." *Information* 10.1 (2019): 18.
7. Yazdavar, A. H., et al. "Mental health analysis via social media data." In *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, IEEE, 2018.
8. Ouyang, G., Dey, D. K., and Zhang, P. "Clique-based method for social network clustering." *Journal of Classification* 37.1 (2020): 254-274.
9. Xu, G., Dong, C., and Meng, L. "Research on the collaborative innovation relationship of artificial intelligence technology in Yangtze River delta of China: A complex network perspective." *Sustainability* 14, no. 21 (2022): 14002.
10. Ye, Z., Wu, Y., Chen, H., Pan, Y., and Jiang, Q. "A stacking ensemble deep learning model for bitcoin price prediction using Twitter comments on bitcoin." *Mathematics* 10, no. 8 (2022): 1307.
11. Li, C., Yang, Q., Pang, B., Chen, T., Cheng, Q., and Liu, J. "A mixed strategy of higher-order structure for link prediction problem on bipartite graphs." *Mathematics* 9, no. 24 (2021): 3195.
12. Kovács, I. A., Luck, K., Spirohn, K., Wang, Y., Pollis, C., Schlabach, S., Bian, W., et al. "Network-based prediction of protein interactions." *Nature Communications* 10, no. 1 (2019): 1240.
13. Yilmaz, E. A., Balcisoy, S., and Bozkaya, B. "A link prediction-based recommendation system using transactional data." *Scientific Reports* 13.1 (2023): 6905.
14. Bukhori, H. A., and Munir, R. "Inductive link prediction banking fraud detection system using homogeneous graph-based machine learning model." In *2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC)*, IEEE, 2023.

 doi:10.20944/preprints202408.2200.v1

15. Tuninetti, M., et al. "Prediction of new scientific collaborations through multiplex networks." *EPJ Data Science* 10.1 (2021): 25.
16. Lim, M., et al. "Hidden link prediction in criminal networks using the deep reinforcement learning technique." *Computers* 8.1 (2019): 8.
17. Yuliansyah, H., Othman, Z. A., and Bakar, A. A. "Taxonomy of link prediction for social network analysis: a review." *IEEE Access* 8 (2020): 183470-183487.
18. Rai, A. K., Tripathi, S. P., and Yadav, R. K. "A novel similarity-based parameterized method for link prediction." *Chaos, Solitons & Fractals* 175 (2023): 114046.
19. Zareie, A., and Sakellariou, R. "Similarity-based link prediction in social networks using latent relationships between the users." *Scientific Reports* 10.1 (2020): 20137.
20. Fang, Y., Yu, J., Ding, Y., and Lin, X. "Inferring complementary and substitutable products based on knowledge graph reasoning." *Mathematics* 11, no. 22 (2023): 4709.
21. Mutlu, E. C., et al. "Review on learning and extracting graph features for link prediction." *Machine Learning and Knowledge Extraction* 2.4 (2020): 672-704.
22. Turki, T., and Wei, Z. "A link prediction approach to cancer drug sensitivity prediction." *BMC Systems Biology* 11 (2017): 1-14.
23. Wang, W., Wu, L., Huang, Y., Wang, H., and Zhu, R. "Link prediction based on deep convolutional neural network." *Information* 10, no. 5 (2019): 172.
24. Cao, Z., Zhang, Y., Guan, J., and Zhou, S. "Link prediction based on quantum-inspired ant colony optimization." *Scientific Reports* 8, no. 1 (2018): 13389.
25. Ahn, M. W., and Jung, W. S. "Accuracy test for link prediction in terms of similarity index: the case of WS and BA models." *Physica A: Statistical Mechanics and its Applications* 429 (2015): 177-183.
26. Li, X., Li, Q., Wei, W., and Zheng, Z. "Convolution-based graph representation learning from the perspective of high order node similarities." *Mathematics* 10, no. 23 (2022): 4586.
27. Hoffman, M., Steinley, D., and Brusco, M. J. "A note on using the adjusted Rand index for link prediction in networks." *Social Networks* 42 (2015): 72-79.
28. Samad, A., Qadir, M., Nawaz, I., Islam, M. A., and Aleem, M. "A comprehensive survey of link prediction techniques for social network." *EAI Endorsed Transactions on Industrial Networks and Intelligent Systems* 7, no. 23 (2020): e3-e3.
29. Divakaran, A., and Mohan, A. "Temporal link prediction: A survey." *New Generation Computing* 38, no. 1 (2020): 213-258.
30. Cai, L., et al. "Line graph neural networks for link prediction." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.9 (2021): 5103-5113.
31. Chen, J., Wang, X., and Xu, X. "GC-LSTM: Graph convolution embedded LSTM for dynamic network link prediction." *Applied Intelligence* (2022): 1-16.
32. Balvir, U., Raghuwanshi, M. M., and Shobhane, P. D. "Improving social network link prediction with an ensemble of machine learning techniques." *International Journal of Computing and Digital Systems* 16.1 (2024): 1-11.
33. Badiy, M., and Amounas, F. "Embedding-based method for the supervised link prediction in social networks." *International Journal on Recent and Innovation Trends in Computing and Communication* 11.3 (2023): 105-116.
34. Wang, T., Jiao, M., and Wang, X. "Link prediction in complex networks using recursive feature elimination and stacking ensemble learning." *Entropy* 24.8 (2022): 1124.
35. Ayoub, J., Lotfi, D., and Hammouch, A. "Mean received resources meet machine learning algorithms to improve link prediction methods." *Information* 13.1 (2022): 35.
36. Malhotra, D., and Goyal, R. "Supervised-learning link prediction in single layer and multiplex networks." *Machine Learning with Applications* 6 (2021): 100086.
37. Aziz, F., Gul, H., Uddin, I., and Gkoutos, G. V. "Path-based extensions of local link prediction methods for complex networks." *Scientific Reports* 10, no. 1 (2020): 19848.
38. Kumari, A., Behera, R. K., Sahoo, K. S., et al. "Supervised link prediction using structured-based feature extraction in social network." *Concurrency and Computation: Practice and Experience*, vol. 34, 2020, pp. e5839.
39. Ran, Y., Xu, X. K., and Jia, T. "The maximum capability of a topological feature in link prediction." *PNAS Nexus* 3, no. 3 (2024): pgae113.

 doi:10.20944/preprints202408.2200.v1

40. Yang, R., Chen, J., Wang, H., Wang, M., Cui, Z., Leung, V. C. M., and Wang, D. "Adversarial enhanced representation for link prediction in multi-layer networks."
41. Gadár, L., and Abonyi, J. "Explainable prediction of node labels in multilayer networks: a case study of turnover prediction in organizations." *Scientific Reports* 14, no. 1 (2024): 9036.
42. Mamat, N., Othman, M. F., Abdulghafor, R., Alwan, A. A., and Gulzar, Y. "Enhancing image annotation technique of fruit classification using a deep learning approach." *Sustainability* 15, no. 2 (2023): 901.
43. Mienye, I. D., and Sun, Y. "A survey of ensemble learning: Concepts, algorithms, applications, and prospects." *IEEE Access* 10 (2022): 99129-99149.
44. Boutahir, M. K., et al. "Effect of feature selection on the prediction of direct normal irradiance." *Big Data Mining and Analytics* 5.4 (2022): 309-317.