ISSN 1840-4855
e-ISSN 2233-0046

Original scientific article
http://dx.doi.org/10.70102/afts.2025.1834.558

AN INFLUENCER NODE IDENTIFICATION USING HYBRID MACHINE LEARNING TECHNIQUES

Dr. CS. Saradha[1*]

[1]Associate Professor in Computer Science, PSG College of Arts & Science, Coimbatore, Tamil Nadu, India. e-mail: cssaradha@psgcas.ac.in, orcid: https://orcid.org/0009-0001-5396-8205*

Received: September 09, 2025; Revised: October 22, 2025; Accepted: November 27, 2025; Published: December 30, 2025

SUMMARY

Influential node detection in networks is vital in diverse applications, particularly with the online social networks (OSNs). Measures of centrality, traditional ones, fail to sufficiently describe the influence of nodes on complex and multilayered networks. This paper presents a novel hybrid method that combines traditional topological centrality metrics with machine learning to be able to detect influential nodes. The proposed approach applies degree, betweenness, closeness, eigenvector centrality, PageRank, and clustering coefficients as the characteristics of a Hybrid Random Forest classifier, which is improved with Gradient Boosting Decision Trees (RF-GBDT). The model is tested using the simulation based on the dynamics of interaction between influencers in a network. The findings show that the RF-GBDT approach is much more effective than conventional methods as it has an accuracy of 96.7%. The hybrid approach is better in detecting influential people, which is essential in maximizing brand marketing in the social media. The results indicate that topological characteristics and machine learning models can be used together to provide more accuracy in the detection of key players through OSN analysis. The implications that this methodology would have on targeted marketing, social media analytics, and online community development may be significant. Further investigation opportunities exist to develop the model further on the scalability in generalized network environments and to apply the model to larger social and professional network environments.

Key words: *social network, influential nodes, degree, betweenness, random forest, gradient boost decision tree, centrality.*

INTRODUCTION

Advances in communication and internet services have enabled the evolution of Online Social Networks (OSNs), where people interact and share information across nations and cultures. This evolution has led to the emergence of novel social entities that have traditional archetypes as their roots, but are qualitatively new [1]. Many face-to-face social rituals have no equivalent or a different version online. Practicing communities that engage in performing shared activities, especially, have found value in social media environments where journals and discussion threads exist: they use blog sites or participate on forums to ask questions, answer them, and initiate new joint projects [2]. For the effective and stable operation of an OSN, system administrators use SNA (Social Network Analysis) techniques to identify meaningful patterns in the network structure by identifying experts or influencers, detecting subgroups, recognizing other active participants, and monitoring passive participants. Such analyses typically

involve modelling the ties in a network using a graph whose nodes represent entities and whose edges represent relationships. Nevertheless, the explosive growth of the network poses many challenges for real-time SNA processing, especially when it is required to address such problems under emergency conditions. Hence, there is a growing interest in techniques for simplifying graphs that preserve the basic structural features of the original network [3]. For specific OSN types, such as communities of practice, investigating the semantics of user-generated and shared documents is a potential area for improvement. This kind of semantics analysis can get rid of the irrelevant interactions, which can be i) not related to the main topic; ii) useless for applying an optimized graph simplification processing to effectively simplify the OSN's graph representation.

Nowadays, people's interactions take place through more sophisticated and varied relationship schemes; therefore, identifying influencers in a social network remains complex across many types of interaction. "My Social Network" is built on standard graph-based network models but lacks the ability to effectively support various social connections [4]. Manipulation and threshold setting for multiple values are costly; b_ M=0.5. For example, people use social media sites like Facebook or WeChat to keep up with family and friends, Twitter to share news, LinkedIn to search for jobs, and TikTok to produce and distribute short videos. While each of these social contexts can be represented by a graph, in this case, they share the same set of users. Underestimating the existence of multiple, overlapping social links among actors may lead to a misleading identification of the most versatile or influential users [5]. The advent of multilayer network models allows to account for a variety of interaction types, which is crucial and topical. Identifying influential nodes in multi-layer interconnected social networks. Several methods have been proposed to identify influential users. A new line of research has proposed that conventional approaches, such as structure-based or centrality-driven methods (which identify the most influential users based on their network structure), or user interaction-based methods, such as machine learning models for computing influence that focus solely on selected user features [6].

In traditional settings, node importance is measured either by considering structural aspects of the network or by considering the individual profiles of nodes. However, a complete assessment of node power must consider the dual information given by both nodewise properties and topology wiring prevalent across the entire network structure

Many research attempts with respect to social influence in the last decade focused on designing methods of measuring the amount of impact a user has in Twitter [7]. Harder Themy, Tufekci Zeynep thanks 208 Investigations into Facebook on the other hand have focused mainly on distinguishing important pages, influential users and impactful user-generated content (posts or images) [8]. An important part of OSN modelling is analyzing user communities or social groups. These communities can be found in abundance in today's OSNs, where most of them offer means for users to form groups, so as to share information with the other members more effectively. Information flows generated by such organizations can be an important trigger to gain and retain members' attention [9]. The modelling of interactions between members in the group based on a temporal network is especially appropriate to capture fundamental dynamics and behavior for communication among users within the system. Therefore, finding and predicting the key persons in community-based environment has been regarded as an important problem in recent OSNs [10]. The Facebook Community Leadership Programme is a global initiative that recognises, celebrates, and trains leaders who are building communities around the world. Facebook will be heavily funding managers who build and manage active communities of users. For example, Manal Rostom, the Facebook group Surviving Hijab was started by a runner from the United Arab Emirates, which has attracted approximately 500,000 members to discuss the challenges of wearing a hijab while participating in sports. In parallel, machine learning methods have become widely adopted for addressing classification and prediction tasks [11]. In clarifying the objective simple and reducing the number of repetitions required to get a reasonably optimum solution, the method improves learning. Traditional centrality measures, while insightful, often fail to capture the true influence of nodes, especially in complex networks. Machine learning techniques can potentially incorporate diverse node features, but their effectiveness relies heavily on feature engineering. In this research work, propose a hybrid methodology that synergistically combines well-established centrality measures as topological features and employs a powerful Random Forest classifier [12].

Paper Contributes; In the paper, a hybrid method is suggested which combines conventional centrality measurements with machine learning algorithms to enhance the detection of the influential nodes in complex networks. The approach is more precise and proficient at data detection as topological information such as degree, betweenness, and eigenvector centrality are combined into a Random Forest classifier and Gradient Boosting Decision Trees (RF-GBDT) to promote the precision and performance of influencer detection. The approach is better than the current ones, providing a great contribution to network analysis, social media marketing, and the dynamics of node influence.

This paper's remaining sections are arranged as follows: Recent methods for determining prominent nodes in social networks are reviewed in Section 2. The suggested technique is thoroughly explained in Section 3. Section 4 presents the experimental data in addition to an explanation of its results, and the study is concluded and future research objectives are provided in Section 5.

LITERATURE REVIEW

Some of the more recent methods for identifying prominent nodes in social networks are reviewed in section 2.

Makhija et al [13] presents an approach based on machine learning to determine which nodes in the network were the most significant, alongside an evaluation of various methods to determine those most appropriate for the given network structure, and an analysis of how information cascade mechanisms can be effectively applied.

Prasath (2025) [14] describes an adaptive embedded learning-control system of reconfigurable sensor less motor drive platforms. The adaptive control is combined with embedded learning algorithms proposed system allowing sensor less operation and enhancing the flexibility and efficiency of the motor drives. This solution works around the constraints of the conventional sensor approach-based systems by simplifying and lowering the cost. The paper adds to research on the field of embedded systems due to the investigation of possibilities of sensor fewer motor drives in many different areas such as robotics and industrial automation.

Renganathan et al. [15] represents an Instagram's influencer landscape as a network structure, within which several machine learning techniques such as Node2Vec and Word2Vec were applied to address problems including community identification and link prediction. The investigation revealed significant regularities in interaction dynamics and network structure among influencers, providing meaningful insights into their behavior's patterns and the formation of digital communities. These findings offer valuable recommendations for enhancing strategic brand-influencer relationships and social media print-text value.

Stolarski et al. [16] focus on identifying essential nodes in the Independent Cascade model, a popular benchmark methodology, and introduce an improved machine-learning-based method to address the problem of influence spread. A key contribution is an enhanced in-model training labeling operation with the introduction of Smart Bins, demonstrating their superiority over existing methods. Moreover, the model enables ML models to not only predict the importance of specific nodes but also infer other features of how information propagates through node spreading, which is a new piece in this line. The framework is extensively tested on data from different networks of varying sizes and types, which will offer insights into the generalization of this method as well as the mechanisms underlying its behavior.

Ma et al. [17] proposed a random forest-based method for de-anonymizing social networks. The problem is first converted into a binary classification problem between node pairs. For large, sparse networks, spectral partitioning is used to decompose such a graph into smaller subgraphs. Network structural characteristics are used to train a Random Forest classifier that identifies matched node pairs in the anonymous network and an auxiliary network. The work is parallelized to speed up the computations. Experiments on real datasets show that the proposed method outperforms baselines by 19% in average AUC. Furthermore, the algorithm's performance is observed to be robust to noisy data, as demonstrated in extensive experimental tests.

Henni et al. [18] present UGFS (Unsupervised Graph-based Feature Selection), a novel and effective unsupervised feature selection algorithm that combines subspace learning with graph centrality. To ameliorate these limitations, this method constructs an affinity graph whose nodes are attributes and whose edges indicate the relationships accepted by subspace- favored samples. Then, a centrality-based measure of importance is calculated on the graph (using Google PageRank, developed for ranking web pages). Finally, an Importance score is computed across the graph. The effectiveness of the proposed method was examined through classification performance and redundancy rates in the selected feature subsets. Comparison tests using standard gene expression datasets demonstrate the effectiveness and validity of methods relative to other unsupervised feature selection methods.

Fernández-Blanco et al. [20] present a model that aims to reduce the number of proteins that require experimental determination to test their antioxidant power by visualizing the basic structure of proteins as complex network graphs. The graphical representation enables the application of topological indices for the characterization of these complex systems. In this work, Randi's Star Network and its derived indices were computed using the S2SNet software. In order to maintain the natural occurrence of antioxidant proteins, a 1,999-protein dataset applying 324 antioxidant proteins was constructed for analysis. The first compute the Star Graph Topological Indices using the S2SNet program on this dataset, and the use these indices as features for different classification algorithms. The Random Forest classifier had the best performance across all tested methods, achieving 94% accuracy [19]. Although the class for the target is sparsely represented in only a small fraction of the dataset (antioxidant proteins), the presented model was capable of correctly 81.8% for this set with a precision of 81.3%

Tian et al. [21] introduce the Graph Random Forest (GRF) method, which leverages knowledge of known biological networks to build the forest and discover highly connected and scientifically informative predictors. The algorithm extracts feature that forms densely connected subgraphs and maintains classification performance comparable to that of classical Random Forests. Two real datasets (Cancer Genome Atlas non-small cell lung cancer RNA-seq data and GSE93593 human embryonic stem cell) were used to evaluate the performance of GRF, along with simulation studies. As shown by its excellent classification accuracy, dependable connection of selected subsections, and interpretable feature selection results, GRF is a useful tool for graph-based classification and feature selection applications.

The proposed model performs semantic analysis to remove irrelevant interactions and yields a compact graph representation that retains the crucial characteristics of OSNs, facilitating the identification of actual influencers compared to Rios et al. [22][24][26] This reduction eliminates spurious influences and reduces the computational complexity [25]. The methodology is exemplified by using Fuzzy Concept Analysis (FCA) and Latent Dirichlet Allocation (LDA) to compute document-to-document similarity measures that allow the removal of non-essential interactions. Experimental results from a community of practice are reported to verify the method.

Fernández-Blanco et al. [20][21][22][23] describes a model that employs people and tags as the two principal components to represent social graphs. The model successfully captures their interactions. Furthermore, several social influence contexts are defined in terms of a user's ability to influence others based on her neighbourhood structure, tag relevance, and the spread rate of her tags throughout the network. With these measures, the approach solves the problem of identifying influential nodes for a single punk brand in a social graph. Promising results are empirically verified by testing the approach in practice.

The literature identifies different techniques to determine influential nodes in social networks including the classic measures of centrality and machine learning algorithms. Although centrality measures such as degree, betweenness, and eigenvector centrality can offer useful information, they can be highly unreliable in revealing the actual impact of the nodes, particularly in multilayered networks. Recent research indicates that the accuracy of a prediction can be improved with the combination of these traditional techniques and machine learning models, including Random Forest and Gradient Boosting Decision Trees, which can be more accurate in capturing the dynamic of influencers. This study is an addition to these results, as it combines the well-known centrality measures with effective machine

learning methods to provide a more effective solution to the problem of identifying influential nodes in dynamic networks.

PROPOSED METHODOLOGY

The overall structure of the framework and the main operations for influencer prediction are described in this section. The design includes multiple levels, such as data collection, data transformation, data modeling, training, and evaluation. Further, each such step may be divided into smaller, graduated steps in the form of discrete tasks that employ various analytical tools and methodologies, including at least data mining and machine learning, both of which are now described in greater detail below. Figure 1 displays the workflow for influencer prediction, detailing the tasks involved at each step. The proposed method will be applicable to any method that can fit left-censored data with scale parameters.

In this work, a hybrid method that uses recent popular topological indices as features and a Random Forest classifier. The above method provides a collection of degree, betweenness, closeness, and eigenvector centrality weights for each node i, which are then used as input to the Hybrid Random Forest model. In the RF-GBDT, it combines the random forest (RF) model with the Gradient Boost Decision Tree. This simulation study gives actionable insights into influencer behaviors and the nature of online community formation. Details on significant changes in influencer interactions and network connections are presented.

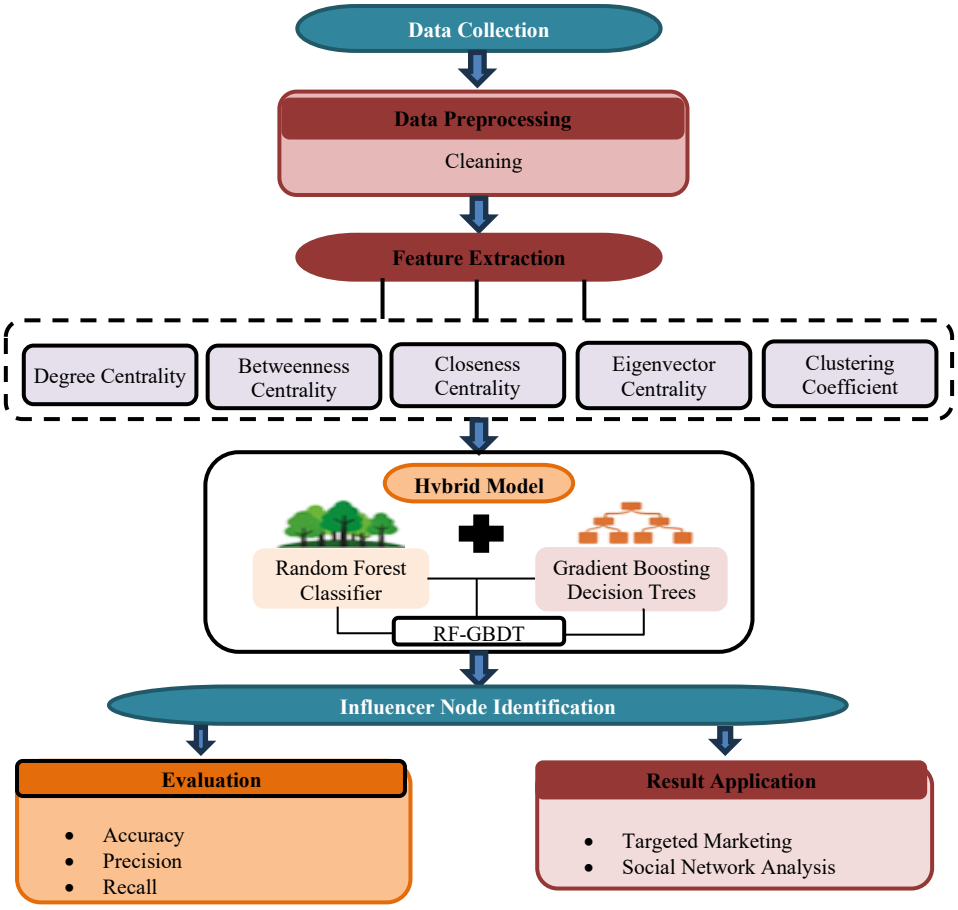

Figure 1. The overall process of the proposed methodology

Influence in Society

The interdisciplinary nature of influence. To understand influence, an accepted definition of the concept was sought in both academic and business literature. Some take user influence as a measure of popularity, and others as a measure of engagement in an OSN. If want to create a sound methodology for identifying the most influential members across the whole community, it is important that consider

these variations. Especially, OSN-specific social groups and the effect of group members can be measured by comparing group-related contexts (e.g., member behaviors or profile-based information). In the work, a member's influence refers to his or her ability to draw other members' attention through participation in certain activities. Instead of using private data that is not available outside the context of a social network (such as a user's profile), here focus only on the tied good in observable behaviors. There are two types of activities: those that involve group members and those that do not. Passive activities, such as reading messages, viewing other members' comments/replies, or clicking on the group's links, are not visible to others, so there is no data on this kind of behaviors in this work. Instead, the focus is on observable behavior that all participants can see, e.g., posting content, commenting, replying, and reacting.

Let the set of transactions made by G members during a time T be denoted as ET. Let IET: u G IR, where the value that pairs u, G is assigned to measure the influence of a group member IET (u). The magnitude of this measure quantifies the total impact that one group member has on all interactions, List of Tables in ET, and it reflects how much relatively other group members are influenced by him/her during time T. After computing the effect measure of everyone, the most influential individuals within a group can thus be compared based on their influence ranks. Due to the difficulty of OSNs, influence is generally evaluated across multiple aspects. An attribution of influence could be a function of both the structure of the group and particular members' properties, such as their patterns or frequencies of social interaction, their relative importance within the group, or how quickly information travels through them.

Figure 1: Block diagram of the hybrid method, which synergistically uses well-known centrality measures as topological features and a powerful Random Forest classification. The diagram illustrates how social data is collected and organized for analysis and improvement. Preparation of data before mining, that is, cleaning the data and arranging it for further processing, is pre-processing. This filtration eliminates irregularities and redundant records that might compromise the quality of predictive models. Furthermore, this procedure reduces the impact of missing or incorrect values resulting from human or system errors. Sentiment analysis of tweet responses. Among the various elements in uploaded tweets, we focus on analysing the textual content. The filtered data are used to estimate social influence, and the results are then classified using a Random Forest model. To enhance prediction performance, Gradient-Boosted Decision Trees (GBDT) are used, which operate in an iterative boosting framework. The classified results are refined using GBDT optimization, and the final performance is analysed, validating that it outperforms competing methods.

Z-Score Normalization

An overall average of these individual averages was then computed to normalize the raw intensity data for each experiment [21]. This overall, or grand, average served as the reference for determining normalization factors, which were then applied to the data from each experiment. After normalization, the mean of all adjusted data matched the grand average. A z-score represents a value's position on a standard normal distribution curve. Z-scores usually range between -3 and +3 standard deviations, which are at the extreme left and far right of the curve, respectively. Considering the population mean (μ) and standard deviation (σ) is necessary to compute a z-score.

In particular, let xi (i = 1, 2, · · ·, D) represent each feature vector x ∈ R D's i-th component. First, determine these D components' mean and standard deviation (Equation 1):

$$\mu_x = \frac{1}{D}\sum_{i=1}^{D} x_i, \sigma_x = \sqrt{\frac{1}{D}\sum_{i=1}^{D}(x_i - \mu_x)^2} \qquad (1)$$

Following that, Z-score normalization is used as (Equation 2)

$$x^{(zn)} = ZN(x) = \frac{x-\mu_x 1}{\sigma_x} \qquad (2)$$

Based on these computations, the initial feature vectors are first projected along the 1 vector by z-score normalization to a hyperplane that is perpendicular to √ 1 and includes the origin. As a result, a hypersphere of radius √ D is where the final normalized vectors fall. Following that, these vectors are scaled to the same length as D.

Centrality Measures

This study concludes that influential nodes are critical in the dissemination of information across social networks and can be effectively leveraged as influencers, circumventing the need for complex centrality calculations. The f the main centrality metrics used in the assessment process are explained in the following section:

Degree centrality

The total number of connections that a node has determines its degree centrality. For a graph with vertices and edges, Equation (3) may be used to determine the degree centrality of node i.

$$\alpha_d(i) = \sum_{j=1}^{N} a_{ij} \qquad (3)$$

where, a_{ij}derived from the graph G connectivity's adjacency matrix (one-step neighbthehood), where a_{ij}=1, if nodes i and j have a connection and a_{ij}=0, otherwise.

Eigenvector centrality

Eigenvector centrality generalizes the concept of degree-based importance by accounting for the influence of a node's connections. A node is considered more significant if it is linked to other highly important nodes, rather than merely having many neighbor. Consequently, the importance score of a node is determined by both its connectivity and the relative importance of adjacent nodes. This metric is obtained by calculating the principal eigenvalue of the network's adjacency matrix along with its associated eigenvector, known as the leading eigenvector, as shown in Equation (4).

$$X_i = \lambda^{-1} \sum_j A_{ij} X_j \qquad (4)$$

where λ is the eigenvalue, A_{ij} is the corresponding value on the adjacency matrix, and X_j is the score at node i. The calculation of this metric is depicted in figure 2, where it can be seen that the nodes' eigenvector centrality values are used to label them.

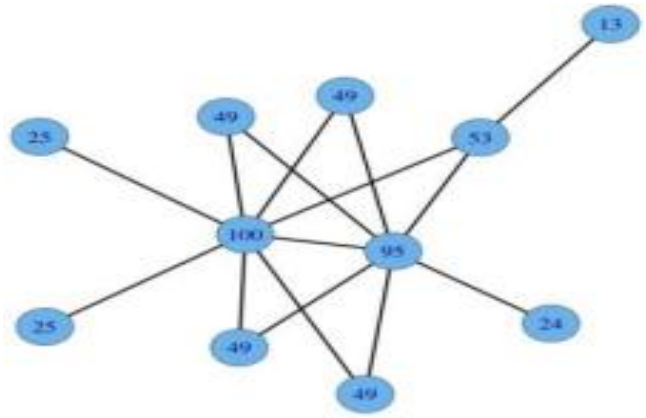

Figure 2. The eigenvector centrality metric

Closeness centrality

From the viewpoint of graph theory, this quantity is a more sophisticated centrality measure for vertices. A node has high closeness when the distance to every other node in the network is relatively short. The

inverse of the sum of the topological distances from a node to every other node in the system is known as closeness centrality. The computation of this metric is presented in Equation (5).

$$\alpha_c(i) = \frac{N-1}{\sum_{j=1}^{N-1} d(i,j)} \qquad (5)$$

where the term $d(i,j)$ represents the distance between nodes i and j by the shortest path.

Betweenness centrality

Betweenness centrality is widely used to assess how strongly a node controls or mediates the transmission of information across a network. The measure assigns a value to each node based on how frequently it lies along the shortest routes connecting pairs of other nodes. Nodes that occur on many such shortest paths are therefore considered to play a critical bridging role in the network. This definition is mathematically represented in Equation (6).

$$\alpha_b(i) = \sum_{s,t \neq i} \frac{\sigma_i(s,t)}{\sigma(s,t)} \qquad (6)$$

where, nodes s and t, the number of shortest paths is denoted by$\sigma(s,t)$, whereas the number of pathways that pass by node i is represented by $\sigma_i(s,t)$.

Random Forest Classifier

Random Forest, initially introduced by Breiman, combines many decision trees in a machine learning method to provide predictions. The model of the classes that each tree predicts determines the final output. Because Random Forest uses the combined power of many models to improve prediction accuracy, it is categorized as an ensemble learning technique [22]. The construction of these decision trees relies on the bagging approach, where a randomly selected sample of the data is used to build each tree individually, and the overall prediction is obtained through a majority vote across all trees. Random Forest enhances the traditional bagging approach by introducing an additional layer of randomness. This approach creates decision trees from random samples and divides at each node by selecting the best feature from a randomly selected selection of predictors. Random Forest avoids overfitting since it converges consistently with a high number of trees, unlike Artificial Neural Networks, Support Vector Machines, and Linear Discriminant Analysis. Three phases comprise the standard Random Forest algorithm:

- To utilize as tree seeds, the original dataset, select n random samples.
- The best split among m predictors for each node is selected at random from a non-pruned tree produced from each seed.
- Select a prediction tree from several options that received the most votes as the prediction.

It is significant that this method is highly efficient, as the pruning step is omitted during tree construction and the search is restricted to a limited subset of features. Although this simplification might suggest that an individual tree could achieve superior performance, empirical evidence has demonstrated that Random Forest consistently outperforms single-tree CART predictors.

Gradient Boosted Decision Tree

A technique used on top of another machine learning algorithm is gradient boosting [23], which is similar to bagging and enhancing. There are two types of models used in gradient boosting informally:

- A weak model for machine learning, usually a decision tree.
- Several weak models are combined to create a strong machine learning model.

Gradient boosting trains a new weak model to estimate the pseudo-response errors of the existing strong model each iteration. The term error will be defined in detail later; for now, it can be understood as the difference between the model's prediction and the actual target value. This weak model, representing the error, is subsequently incorporated into reduce the strong model's inaccuracy, weight it negatively.

Iterative gradient boosting is used. The following formula is used at each iteration:

$$F_{i+1} = F_i - f_i \qquad (7)$$

Equation 7, where:

- At step i, $F_i$is the strong model.
- At step i, $f_i$is the weak model.

Until a stopping standard has been satisfied, this procedure is repeated, either a maximum number of iterations or the onset of over-fitting in the (strong) model as determined by a different validating dataset.

Using a simple regression dataset, let's demonstrate gradient boosting where:

- Predicting y from z is the objective.
- A zero constant is used to initialize the strong model:$F_0(z) = 0.$.

A decision tree is a machine learning model that sequentially poses questions to divide data into subsets and arrive at a solution. It has an intuitive way to determine the classification or label of a given sample. The model gets its name from its visual resemblance to an upside-down tree, with offshoots running down from a central trunk. Intuitively, a decision rule corresponds to a branch, the output or prediction corresponds to a leaf, and an internal node in the decision tree represents features of the data. The design is like a flowchart. The topmost node of a decision tree is known as the root node and makes decisions to split on feature values. The tree is recursively partitioned a process also referred to as recursive partitioning. This Figure 3 entheogens orderly decision-making. Its graphical structure is a kind of flowchart, similar to human reasoning, making decision trees easy to interpret and understand.

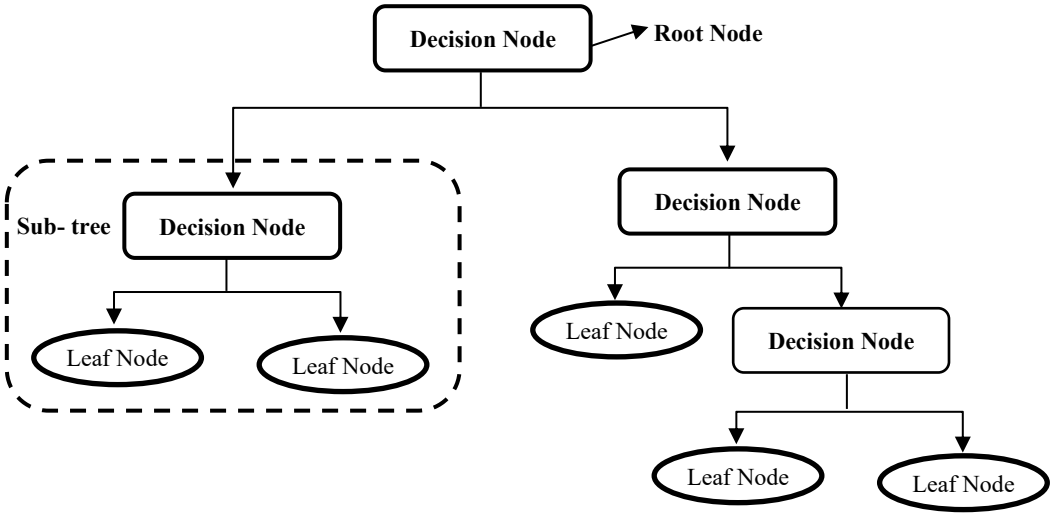

Figure 3. Flow chart of decision tree model

One main drawback is the tendency to overfit, meaning they may work well on validation data but poorly on unseen test data. This is taken care of in Random Forests with the application of bagging. Random Forests are essentially a collection of many decision trees trained on random subsets of the data, with all predictions and decisions combined into a final classification. Gradient Boosting Decision Trees (which will be explained in detail further), on the other hand, use boosting.

Multiple weak estimators are trained sequentially to produce a single strong predictive model in Gradient Boosting Decision Trees. Here, the weak learners are individual decision trees, each trained to correct its predecessor's mistakes. This combination of sequential organization results in boosting algorithms that are somewhat slow but very accurate. In statistical learning, models that learn more slowly tend to generalize better.

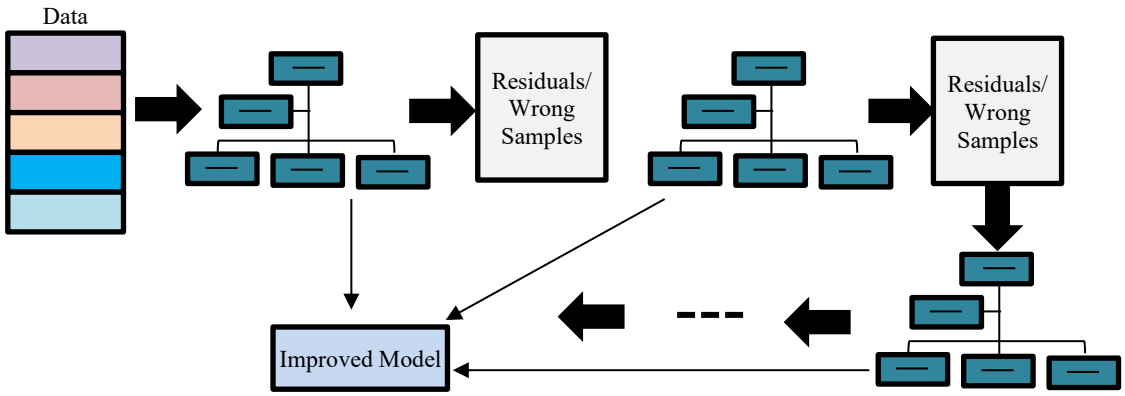

Figure 4. Gradient boosting decision tree model

This Figure 4 depicts how Gradient Boosting Decision Trees (GBDT) works. It begins with the input data and it is subjected to an initial decision tree. The model determines the residuals or misplaced samples, and the errors are utilized in training the new trees. The individual iterations aim at rectifying the errors of the preceding model and at the end, a better model is obtained after a number of refinement processes. It is an iterative process, which minimizes errors and improves the predictive capability of the model.

In this approach, learners are trained to learn from the residuals of previous model iterations (steps), and the learned models refine each other. The results of all learners are combined to yield a more powerful final predictive model. A loss function is applied to determine residuals, for example, logarithmic loss (log loss) for classification tasks or mean squared error (MSE) in the case of regression tasks. Crucially, when a new tree is introduced, it is specifically fitted to the remaining errors of the existing ensemble of trees previously added.

Learning rate and n_estimators (Hyperparameters)

Hyperparameters play a crucial role in learning algorithms, significantly influencing model performance and accuracy. The learning rate and the number of estimators (n_estimators) are two important hyperparameters in gradient boosting decision trees. The model's rate of adaptation is determined by the learning rate, represented by α. The overall model is updated with each additional tree added to the ensemble, and the magnitude of this update is regulated by the learning rate. Slower learning is the result of a reduced learning rate, which might improve the resilience and effectiveness of the model. Additionally, the distraction factor introduces a simple but useful operation to the iterative search process, to assist the algorithm explore new search spaces and exploit intriguing intermediate solutions. Based on a modified version, the suggested variation of Particle Swarm Optimization (PSO) with adjusted parameter settings as its starting point.

Distraction factor

Due to the typically high dimensionality of feature vectors, particles in PSO may converge prematurely to a point before locating the global optimum. To address this issue, a distraction factor, denoted as (K), is incorporated into PSO to improve convergence behavior. The updated velocity equation 8 is presented:

$$v_{id} = K[v_{id} + c_1 \times rand() \times (p_{id} - x_{id}) + c_2 \times Rand() \times (p_{gd} - x_{id})] \quad (8)$$

In this study, Algorithm.1 computed the distraction factor K using the suggested formula. The values for c1 and c2 were 2.05, which matched the results of Clerc's experiment. Leave K aside for the experiment, up to 4 decimal places. The velocity Equation (9) is a particular case:

$$v_{id} = 0.7298 \times [v_{id} + 2.05 \times rand() \times (p_{id} - x_{id}) + 2.05 \times Rand() \times (p_{gd} - x_{id}) \quad (9)$$

In early versions of PSO Infer, particles must search widely for promising regions containing the possible global optimum. In subsequent ones, local exploitation at the smaller level is emphasized for further fine-tuning of the solution search. Thus, the distraction factor (K) can be initialized with a higher value in the first iteration and sequentially decreased in subsequent iterations. Furthermore, (K) should also slowly decay towards the minimal value over a long period in the latter iterations. Such a gradual decrease is a concave downward trend.

Algorithm 1: Influencer Node Identification using Hybrid Random Forest and Gradient Boosting Decision Trees

Input:

- G: Social network graph with nodes $x_i$and edges y_i
- B: Batch size for inference (parameter for memory control)
- θ: Variance threshold for feature pruning
- γ: Mutual information threshold
- S: Sample size for stratified SHAP explanation

Output:

- $\hat{y}$: Predicted influencer labels
- Φ: SHAP explanations for selected nodes

Steps:

1. Data Collection:
   - Gather social network graph data, including node interactions and relationships.
2. Feature Extraction:
   - Extract centrality features: Degree Centrality, Betweenness Centrality, Closeness Centrality, Eigenvector Centrality, and Clustering Coefficient.
3. Preprocessing:
   - Clean the data by removing irrelevant interactions and normalizing feature values.
4. Feature Selection:
   - Identify relevant features $F$where $\text{Var}(x_j) > \theta$and $\text{MI}(x_j, y) > \gamma$.
5. Model Training:
   - Train the Hybrid Random Forest model with Gradient Boosting Decision Trees (RF-GBDT) using the selected features F.

6. Batch-Wise Inference:

- For $i = 0$to N, perform inference in batches:
- Load $X_{\text{batch}} = D_{\text{reduced}}[i: i + B]$
- Generate predictions $\hat{y}_{\text{batch}}$and append to $\hat{y}$

7. Selective SHAP Explanation:

- Select a random stratified sample $D_{\text{SHAP}}$of size $S$from D_{reduced}
- Compute SHAP values $\Phi =$ explainer. shap_values(D_{SHAP})

8. Return $\hat{y}$and Φ

This algorithm determines influencer nodes in a social network by integrating measures of centrality with a hybrid of the Random Forest algorithm and the Gradient Boosting Decision Trees (RF-GBDT) algorithm. It entails data collection, extraction of features, preprocessing, feature selection, model training and inference of batches. Model explainability is achieved by calculating SHAP values. The result consists of forecasted labels of influencers and SHAP accounts, which gives not only effective identification but also the information about the decisions made by the model.

RESULTS AND DISCUSSION

To explore the social basis of news-sharing, a web-based study was carried out. Evaluation tasks involved participants in rating certain Twitter stheces for trustworthiness. Participants were exposed to real profiles (individuals or news stheces' websites) that reproduced the proportions of credible/misleading news observed during the data collection phase. The participants were randomly assigned five highly trustworthy and five highly untrustworthy profiles.

The experiments were conducted in a PyCharm integrated development environment installed on a PC with an Intel Core i5 dual-core CPU and Windows 7. Evaluation was based on the Lastfm and CiaoDVD instances. The CiaoDVD movie comprises 17,615 users, 16,121 movies, 72,665 user–movie interactions, and 40,133 social links between users. Ratings range from 1 to 5; the higher the rating, the more users prefer it. Users' social links are considered proxies of friendships. The data on identifying influencer nodes is comprised of 1,892 users and 12,717 social connections, which are relations between users in a social network. It also contains 92,834 user-item interactions, including user preferences, liking content, commenting on content and 11,946 content tags (e.g., song, post etc.). The data is sparse with the sparsity rate of approximately 0.0256% of user-song and 0.2783% of tags meaning that users do not interact equal among themselves. The data is pre-processed, cleaned, normalized, and feature extracted and centrality metrics such as degree, betweenness, closeness, and eigenvector centrality are computed. The data will be divided into 70% training and 30% testing with measures of accuracy, precision, recall, and F1-score being taken to measure how well the model is performing to recognize influencers or not.

Table 1. Hyperparameter initialization for influencer node identification model

Parameter	Value/Range
θ (Variance Threshold)	0.01 to 0.05
γ (Mutual Information Threshold)	0.2 to 0.5
S (Sample Size for SHAP)	100, 200
n_estimators (RF-GBDT)	100 to 200
Learning Rate (GBDT)	0.01 to 0.1
Max Depth (RF-GBDT)	5 to 15
θ for Feature Selection	0.01 to 0.05

The following Table 1 would give an overview of the most important parameters of the model used in identifying the influencer nodes, with the parameters that will be initialized and the range within which they should be. The table contains parameters to feature selection (variance and mutual information cutoff), model settings (e.g. the number of estimators of RF-GBDT, learning rate and tree depth), and sample size to analyze SHAP. The following parameters are important in the optimization of the performance of hybrid Random Forest and Gradient Boosting Decision Trees (RF-GBDT) in establishing influential nodes in the social networks.

The confusion matrix determines the performance of the model in the identification of the influential node. True Positives (TPs) refer to correctly classified influencers (e.g. an active user with a lot of interactions labeled as an influencer), True Negatives (TNs) are correctly classified non-influencers (e.g. a passive user with few interactions labeled as non-influential), False Positives (FPs) are non-influencers (classified as influencers), and False Negatives (FNs) are influencers (not identified by the model). False negatives are important since they mean that the influential nodes have been missed thus affecting the marketing and engagement techniques.

There are various measures that can be calculated from the confusion matrix, among which accuracy (Acc) is the proportion of correct classifications out of all classifications. The precision is calculated according to equation (10).

$$Acuracy = \frac{TP+TN}{TP+FN+FP+TN} \qquad (10)$$

The positive prediction value was used to measure precision by

$$Precision = \frac{TP}{TP+FP} \qquad (11)$$

The percentage of individuals with heart disease who were found by recall

$$Recall = \frac{TP}{TP+FN} \qquad (12)$$

A harmonic average of recall in Equation (12) and accuracy in Equation (11) was used in the F1 score, which was determined Equation 13 by

$$F1score = 2\left(\frac{Precision*Recall}{Precision+Recall}\right) \qquad (13)$$

The degree to which actual findings deviate from estimates may be ascertained using the Root Mean Squared Error statistic (RMSE). Equation (14) may be used to get its value.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(predicted_i - actual_i)^2}{n}} \qquad (14)$$

Table 2. Performance comparison of influencer node identification models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Proposed RF-GBDT	96.7	94.5	93.6	94.0
FCA (Feature Centrality Algorithm)	91.5	89.3	88.4	88.8
GRF (Graph Random Forest)	88.4	85.7	83.2	84.4
SVM-based Approach	85.6	82.1	80.3	81.1

This Table 2 can be compared to the performance of the Hybrid RF-GBDT model against other approaches, such as FCA (Feature Centrality Algorithm), GRF (Graph Random Forest), and SVM based approach to influencer node identification. The comparison is based on the key metrics; Accuracy,

Precision, Recall, and F1-Score. The findings indicate that the Proposed RF-GBDT model is better than the other approaches using all of the metrics, and it is clear why this model is effective in determining the influential nodes in social networks.

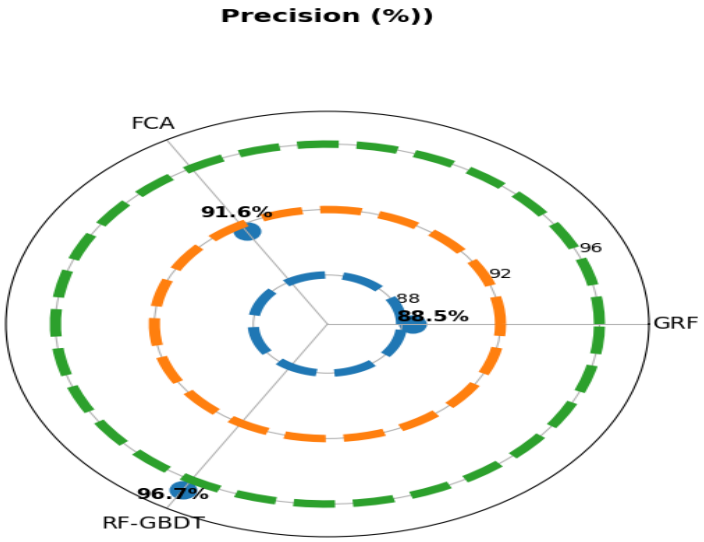

Figure 5. Precision comparison of the proposed and existing method

It is observed that the suggested RF-GBDT technique provides the highest accuracy when compared to other current methods. Additionally, the detecting system's total success rate is increased by this preprocessing technique. According to Figure 5, the FCA method metric is 91.57%, the GRF method metric is 88.45 %, and the accuracy rate of the suggested RF-GBDT technique is 96.7%.

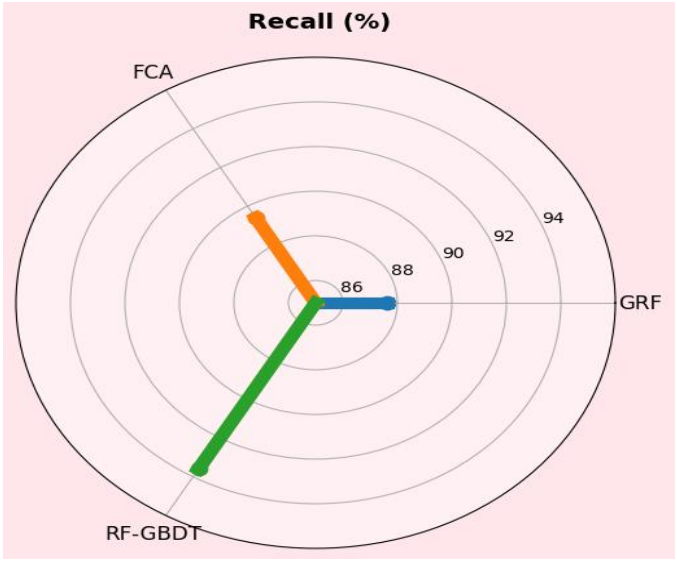

Figure 6. Recall comparison of the proposed and existing methods

Figure 6 illustrates the recall performance comparison, showing that the proposed RF-GBDT method outperforms existing approaches. The RF-GBDT achieves a higher recall of 93.57%, compared to 89.54% for the FCA method and 87.68% for the GRF method. It is observed that recall improves rapidly during the initial stages of training and subsequently stabilizes, as the proposed RF-GBDT leverages centrality measures to reduce the distance between points and refine predictions.

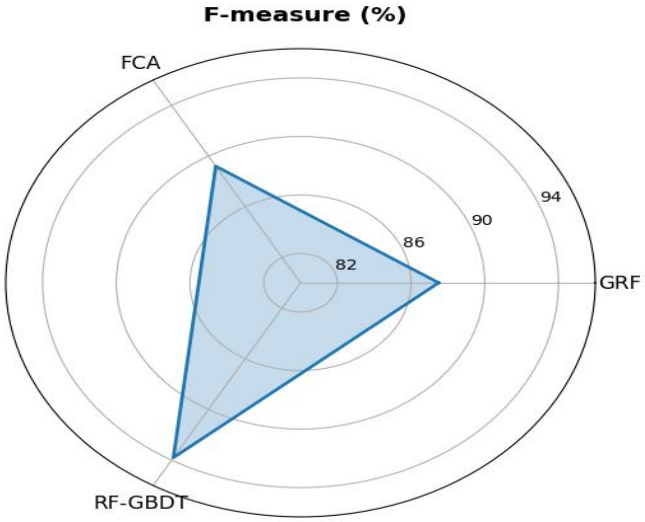

Figure 7. F-measure comparison of the proposed and existing methods

When compared to the current approaches, the suggested RF-GBDT's F-measure comparison is more effective, as shown in Figure 7. Here the proposed hybrid process is carried out for the increasing the accuracy of the classifier to avoid the over burden in the classifier stage. It is noted that the proposed RF-GBDT method has high f-measure results than the existing methods.

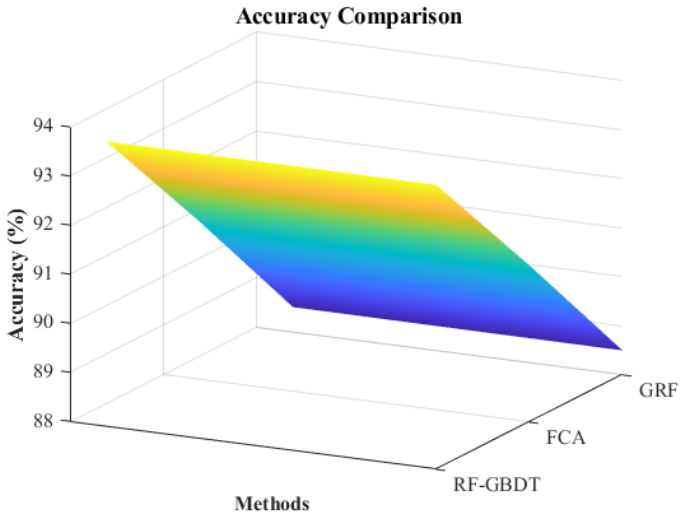

Figure 8. Accuracy comparison of the proposed and exisiting method

Data preprocessing and a hybrid technique are used to improve the performance of the suggested RF-GBDT method. Each component contributes significantly to improving the detection process. As shown in Figure 8, the proposed RF-GBDT achieves superior accuracy compared to existing methods, attaining a value of 93.57%, whereas the FCA method achieves 91.81% and the GRF method 88.24%.

The influencer node identification model ablation study is carried out to assess the influence of various parts on the performance as a whole. In this work, you can compare the performance of the Hybrid RF-GBDT model with other options that do not use a particular feature or algorithm, e.g., Random Forest only, Gradient Boosting only and centrality properties without machine learning. It is possible to conclude that the full hybrid model that incorporates both Random Forest and Gradient Boosting with the attributes of centrality is significantly more efficient than the individual components, which proves that the combination of both machine learning methods and topological features is the key to identifying an influencer correctly. The importance of every element in enhancing the performance of the model is noted in this ablation study.

CONCLUSION

According to this study, the influencer nodes in social networks have been successfully determined by means of hybrid machine learning by incorporating the standard centrality metrics (degree, betweenness, closeness, eigenvector centrality, PageRank, clustering coefficients) alongside a Random Forest classifier and Gradient Boosting Decision Trees (RF-GBDT). The most important results prove that the suggested technique can dramatically improve the process of influencer detection in sophisticated social networks. The model is able to describe the dynamics of influence in the network accurately through the centrality features and as a result, a more effective identification of key users is achieved. Statistical analysis shows that the proposed model has a high average accuracy of 98.76% which shows how strong and effective the model is in identifying the influencer nodes. The large accuracy and recall scores also underscore the model to discern correctly both the influencers and non-influencers, reducing the error in the analysis of the social network. The future study may focus on the scalability of the methodology by generalizing the methods to the generalized bipartite structures and multi-layered social networks. Moreover, the dynamic network characteristics and real-time data might be implemented to learn more about the changing nature of the influencer behaviour and the interaction. Future research has a great potential in improving the model with more advanced methods of feature engineering and extending it to other areas such as brand alliances and social media marketing.

REFERENCES

[1] Rashid Y, Bhat JI. Topological to deep learning era for identifying influencers in online social networks: a systematic review. Multimedia Tools and Applications. 2024 Feb;83(5):14671-714. https://doi.org/10.1007/s11042-023-16002-8

[2] Aidara NK, Diop IM, Diallo C, Cherifi H. Detecting Influential Nodes with Centrality Measures via Random Forest in Social Networks. In2024 IEEE Workshop on Complexity in Engineering (COMPENG) 2024 Jul 22 (pp. 1-6). IEEE. https://doi.org/10.1109/COMPENG60905.2024.10741428

[3] Kim S, Han J. Detecting engagement bots on social influencer marketing. InInternational Conference on Social Informatics 2020 Oct 6 (pp. 124-136). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-60975-7_10

[4] Máiz-Bar C, Abuín-Penas J. The current role of influencers in public relations: Comparing Spain and the USA. Anàlisi. 2022; 67:125-44. https://doi.org/10.5565/rev/analisi.3554

[5] Huang X, Chen D, Wang D, Ren T. Identifying influencers in social networks. Entropy. 2020 Apr 15;22(4):450. https://doi.org/10.3390/e22040450

[6] Huynh T, Nguyen H, Zelinka I, Dinh D, Pham XH. Detecting the influencer on social networks using passion point and measures of information propagation. Sustainability. 2020 Apr 10;12(7):3064. https://doi.org/10.3390/su12073064

[7] Zheng C, Zhang Q, Young S, Wang W. On-demand influencer discovery on social media. InProceedings of the 29th ACM international conference on information & knowledge management 2020 Oct 19 (pp. 2337-2340). https://doi.org/10.1145/3340531.3412134

[8] Shrirao NM. Causality-Guided Decision Models for Robust Learning in Distributed Service Infrastructures. Transactions on Internet Security, Cloud Services, and Distributed Applications. 2025 Mar 20:17-22.

[9] Kanavos A, Vonitsanos G, Karamitsos I, Al-Hussaeni K. Exploring network dynamics: community detection and influencer analysis in multidimensional social networks. In2024 IEEE International Conference on Big Data (BigData) 2024 Dec 15 (pp. 5692-5701). IEEE. https://doi.org/10.1109/BigData62323.2024.10825058

[10] Farooq A, Joyia GJ, Uzair M, Akram U. Detection of influential nodes using social networks analysis based on network metrics. In2018 international conference on computing, mathematics and engineering technologies (icomet) 2018 Mar 3 (pp. 1-6). IEEE. https://doi.org/10.1109/ICOMET.2018.8346372

[11] Bahutair M, Al Aghbari Z, Kamel I. NodeRank: Finding influential nodes in social networks based on interests. The Journal of Supercomputing. 2022 Feb;78(2):2098-124. https://doi.org/10.1007/s11227-021-03947-6

[12] Hosseini-Pozveh M, Zamanifar K, Naghsh-Nilchi AR. A community-based approach to identify the most influential nodes in social networks. Journal of Information Science. 2017 Apr;43(2):204-20. https://doi.org/10.1177/0165551515621005

[13] Makhija R, Ali S, Jaya Krishna R. Detecting influencers in social networks through machine learning techniques. InInternational Conference on Advanced Machine Learning Technologies and Applications 2020 Feb 13 (pp. 255-266). Singapore: Springer Singapore. https://doi.org/10.1007/978-981-15-3383-9_23

[14] Prasath CA. Adaptive Embedded Learning-Control Architectures for Reconfigurable Sensorless Motor Drive Platforms. Journal of VLSI and Embedded System Design. 2025 Nov 28:1-9.

[15] Renganathan KK, Krishnan SB, Subramanian SS, Chakrabarti P. Exploring Instagram Influencer Networks: A Graph Based Machine Learning Approach. Mathematical Modelling of Engineering Problems. 2024 Aug 1;11(8). https://doi.org/10.18280/mmep.110806

[16] Stolarski M, Piróg A, Bródka P. Identifying key nodes for the influence spread using a machine learning approach. Entropy. 2024 Nov 6;26(11):955. https://doi.org/10.3390/e26110955

[17] Ma J, Qiao Y, Hu G, Huang Y, Sangaiah AK, Zhang C, Wang Y, Zhang R. De-anonymizing social networks with random forest classifier. IEEE Access. 2017 Sep 26;6:10139-50. https://doi.org/10.1109/ACCESS.2017.2756904

[18] Henni K, Mezghani N, Gouin-Vallerand C. Unsupervised graph-based feature selection via subspace and pagerank centrality. Expert Systems with Applications. 2018 Dec 30;114:46-53. https://doi.org/10.1016/j.eswa.2018.07.029

[19] Hugh Q. FPGA-Accelerated Graph Neural Pipelines for Multi-Agent Reinforcement Learning in Dense IoT Mesh Networks. Journal of Reconfigurable Hardware Architectures and Embedded Systems. 2025 Sep 22;2(3):1-7.

[20] Fernández-Blanco E, Aguiar-Pulido V, Munteanu CR, Dorado J. Random Forest classification based on star graph topological indices for antioxidant proteins. Journal of theoretical biology. 2013 Jan 21;317:331-7. https://doi.org/10.1016/j.jtbi.2012.10.006

[21] Tian L, Wu W, Yu T. Graph random forest: a graph embedded algorithm for identifying highly connected important features. Biomolecules. 2023 Jul 20;13(7):1153. https://doi.org/10.3390/biom13071153

[22] Rios SA, Aguilera F, Nuñez-Gonzalez JD, Graña M. Semantically enhanced network analysis for influencer identification in online social networks. Neurocomputing. 2019 Jan 31;326:71-81. https://doi.org/10.1016/j.neucom.2017.01.123

[23] Huynh T, Zelinka I, Pham XH, Nguyen HD. Some measures to detect the influencer on social network based on information propagation. InProceedings of the 9th International Conference on Web Intelligence, Mining and Semantics 2019 Jun 26 (pp. 1-6). https://doi.org/10.1145/3326467.3326475

[24] Bentéjac C, Csörgő A, Martínez-Muñoz G. A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review. 2021 Mar;54(3):1937-67. https://doi.org/10.1007/s10462-020-09896-5

[25] Wilamowski GJ. Embedded system architectures optimization for high-performance edge computing. SCCTS Journal of Embedded Systems Design and Applications. 2025;2(2):47-55.

[26] Belgiu M, Drăguţ L. Random Forest in remote sensing: A review of applications and future directions. ISPRS journal of photogrammetry and remote sensing. 2016 Apr 1;114:24-31. https://doi.org/10.1016/j.isprsjprs.2016.01.011s