International Journal of Modern Computation, Information, and Communication Technology
2022;5(1):1-10.
ISSN: 2581-5954

Deep learning algorithms for analyzing social network influencers

Deepak Maurya[1], Sunita Yadav[1*], Jay Kant Pratap Singh Yadav[1]

[1]*Department of Computer Science & Engineering, Ajay Kumar Garg Engineering College, Uttar Pradesh, India.*

[*]*Corresponding author: yadavsunita@akgec.ac.in*

Abstract. Recognizing influential users is critical in large networks because of its diverse fields. Conventional centrality techniques are often based on topographical network architectures, with different centrality methods taking into account various structural features associated to functional relevance. In several contexts, however, there is always a complicated and nonlinear relationship between a node's functional importance and its many properties, such as local location, worldwide location, and so on, which is difficult to characterize by a single centrality. This study offers a method that is based on machine learning to quantify the relevance of vertices in the scenarios of propagation in order to address this problem. When it comes to supervised learning models, such as KNN, DT, RF, SVM, and CNN ELM, distinct supervised predictors are constructed in order to identify the nodes in complex social networks that have the greatest influence on the network as a whole. In terms of accuracy, sensitivity, and specificity, as well as the F1-Score measure, the CNN ELM model that was developed performed better than alternative learning algorithms. The average F1-score that CNN ELM managed to attain was 94.2%, and its detection exactness was roughly 98.5%.

Keywords: Centrality, Machine Learning, Influential Nodes, Complex Networks, CNN_ELM

INTRODUCTION

In the research field, recognizing prominent nodes in complicated systems has gotten a lot of attention [1]. In recent years, a variety of methods for detecting important nodes within complex networks have been up as potential solutions. The capacity of a node to disseminate information offers novel perspectives on a variety of issues, including the regulation of the exchange of messages and accusations within social networks; the ranking of the reputations of scientists; and the identification of social leaders, amongst other topics [2]. In the 1950s, Shimbel proposed the use of Stress Centrality as an early method for locating notable nodes. [3] He made the suggestion that the whole list of the shortest routes that go through a node should be utilized to assess the node's degree of centrality. Degree centrality is a straightforward and efficient method for identifying the significance of individual nodes, but it is blind to the general architecture of the network [4]. The relevance of a node's neighbors is taken into account by Eigenvector Centrality. Between's Centrality and Closeness Centrality require prior knowledge of all network topology information and so cannot be used to networks [5]. Real social networks, particularly large-scale social networks, frequently feature a visible community structure with strongly related nodes. However, it's worth noting that the nodes in different communities are frequently only loosely connected [6]. Because communities are typically much smaller than the total network, searching for seed nodes within each one can significantly reduce computational overhead [7]. The CGA (Community-based Greedy Algorithm) partitioned the network into numerous communities and found seed nodes in each community using the Mix Greedy algorithm [8]. The CGA, on the other hand, overlooks the bridge or hub nodes that connect several communities. Because bridge or hub nodes can disseminate influence across several communities, limiting the search limit of these nodes in the community is inappropriate.

Received: 20.12.2021 Revised: 18.01.2022 Accepted: 25.01.2022

The author used source nodes from the center and boundary of each community to limit the loss of influence spread caused by community division in order to solve this problem [9]. Nonetheless, in real social networks, node influence follows a power law distribution, and many nodes with minimal impact find it challenging to become seed nodes [10]. As a result, by utilizing the influence distribution features of nodes, many needless Monte Carlo simulation steps can be eliminated, improving the algorithm's efficiency while assuring influence dispersion. The limitations of these traditional algorithms are less efficient in large and complex networks such as medical field, big-data, HPC, industries, and other applications. To address this, it develops these issues; this developed the most trending supervised learning models which is pre-trained with historical data in order to detect the most significant node in the complex network based on its centrality, and architecture features.

1.1. Significant Contribution

In this paper, distinct categories of machine learning networks are adopted and modified as binary classifier in order to detect the node is most influential or not among the given complex network nodes. The primary objective in detecting these influential nodes is to avoid the catastrophic situation in the dynamic environment.

- The social networks such as “WhatsApp, Facebook, and Instagram”, are studied widely in order to understand its working structure in terms of number of users, its transmission bandwidth, and its communication network etc.
- Then the significant features for detecting the influential nodes in these social networks are extracted using the CNN algorithm which is feed-forwarded into the training network in the next phase.
- Five different labelled models such as “KNN, DT, RF, SVM, and CNN_ELM” are trained with the extracted features in order to identify the most significant or influential node in these complex networks automatically.
- The proposed models are evaluated with the public dataset available in the Kaggle network which is utilized for both training and testing purpose.
- Finally, the proposed CNN is combined with ELM algorithm in order to obtain more accurate results with less time complexity. The proposed CNN_ELM achieved nearly 98.2% in accuracy, 94.3% F1 score when compared to other labeled networks

Outline

The organization of the paper: Section 2 details the literature survey; Section 3 depicts the proposed labeled networks along with the Boosted CNN model. Then the experimental results are described in Section 4 with its results and discussion. The conclusion is described in Section 5.

LITERATURE SURVEY

Real-time applications, such as Twitter, social networks, G-maps, and many more, are often organized as network graphs (i.e. nodes and links). In today's world, endeavoring to become influential at nodes in networks as complicated as these is a significant difficulty. To overcome this issue, researchers are focusing on developing learning models to identify the most significant nodes within each network.

A neural network graph framework has been given the moniker "Dr. BC," to recognize the between-node centrality (BC) among the network's nodes. Encoders and decoders are the two key components that comprise this conventional approach. It is possible to capture the structural information of each node by using something called a neighborhood encoder. The relative rank of each node is determined by the MLP decoder based on the BC value, and it is prioritized during program execution. This learning approach incorporates a small-network graph that is directly applied on a network that exists in the actual world [11]. The k-shell decomposition technique was employed in the proposed method to separate nodes and construct clusters, a novel clustering approach for identifying prominent nodes in social networks. A strategy called extended clustering coefficient ranking was developed with the goal of grouping local nodes based on their degree of similarity to one another and their degree of correlation with each other. A node with a low correlation ratio is characterized as highly influential within a social network due to its ability to transmit messages to various regions of the network. The suggested model suffers

from the drawback that its temporal complexity becomes difficult to manage when the network size is extremely large [12].

One way to find the most important node in a social network is to use a reinforcement learning algorithm like Q-RNN with a seed selection strategy. The deep learning neural network based influence maximization (DISCO) algorithm incorporates a number of states to capture the structural information of nodes. On the basis of the rewards obtained at each node, important nodes have been identified. The time-consuming and expensive diffusion sampling step may be skipped using DISCO [13].

The approaches that are based on the semantic graph for keyword abstraction are unsupervised and have as their primary goal the construction of a network of words. Following this step, the nodes in the network will be ordered according to centrality metrics. The approach that has been presented involves the extraction of features from information on Twitter, as well as an algorithm for the extraction of keywords in terms of between centrality, eigenvector, degree, and so on. The page ranking method is used to sort the influential nodes as they are being discovered at runtime [14]. The extracted characteristics include semantic features as well as node attributes, and their purpose is to identify the top ten most influential nodes in the Twitter network.

By utilizing inclination connection inspection and the arbitrary walk process, a new essential centrality measurements and complete calculation has been suggested and given the name PARW-Rank. This calculation is intended for use in determining how hub influences are distributed. The inclination connection that exists between each hub match in an arrangement is investigated in order to construct the midway inclination chart (PPG) [15]. This is done for each critical degree.

A system both visualizes the intricate social structures and locates compelling hotspots in the arrangement of the data. One of the most used algorithms, known as the K-truss composition technique, is one that plays a significant role in the visualization and analysis of social networks as well as the identification of persuasive hubs [16]. The normal exhibitions of machine learning classifiers were much way better than that of conventional centrality strategies on assessing hubs proliferation capacities; the machine learning classifiers would have way better exhibitions than centrality strategies in case it was prepared in a small-scale organize and tried in a large-scale organize, but not way better or indeed more regrettable exhibitions than centrality strategies in case it was prepared in a large-scale organization; the machine learning classifiers would have way better exhibitions than centrality strategies in case [17].

Object recognition based on empirical wavelet transform approach in which the classification is complete with K- nearest neighbor [18]. Switched Inductor Quasi Z Source Inverter utilizes the reactive factor in the main circuit that raises the voltage and minimizes the voltage ripple. The solar photovoltaic has applied with highest power point traction technique to remove the maximum power from the photovoltaic method [19]. Non Linear Feedback shift registers are utilized over Automatic Test Pattern to recognize the Stuck at faults that provides less energy equate to the predictable approach with great fault coverage. K-Means clustering based segmentation approach applying a ship can be segmented and recognized which object seen over the sea. Peak Signal to Noise Ratio values are measured for receiving the function process this approach. A Convolutional Neural Network is used to identify spectral color in order to diagnose glaucoma. Here, color modules are divided as red, green, and blue. Next, the Green channel is applied for the image analysis and detection.

PROPOSED METHODOLOGY

In order to determine which nodes in popular social complex networks are the most influential, the authors of this research build labeled networks that are networks that have already been trained. The major objective is to minimize the amount of time spent searching for the node and to automatically identify it among the "N" number of other nodes that make up the complicated network. The suggested labeled algorithm's high-level structure is shown in figure 1, which also includes its name.

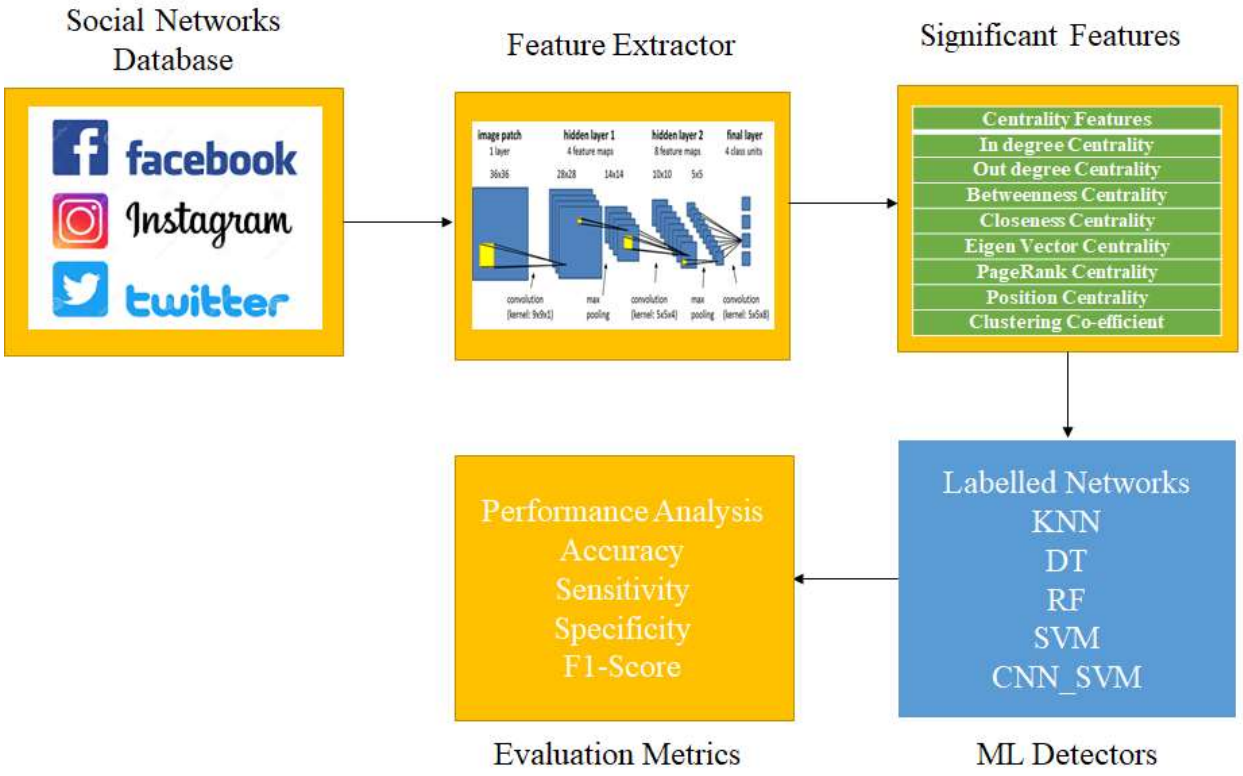

FIGURE 1. Overview Structure of Proposed Framework

Feature Extraction using CNN

Convolutional neural networks (CNN) are used to automatically extract, and then SVM is used to further classify opinions. The information is standardized before being applied in two layers for further processing. For bag of sentiments extraction, the word2vec calculation is adopted in the first layer.

Convolutional Neural Network

A convolutional neural network (CNN) is comprised of many layers of convolutional processing, and each convolutional layer may be followed by a pooling layer that calculates the average or maximum of a window of neurons. This layer makes the output feature maps simpler while also achieving translation invariance and reducing the amount of overfitting that occurs. Fully connected (FC) or dense layers are the last layers of a convolutional neural network (CNN). These layers link all of the neurons from the preceding layers to ensure that the complex characteristics collected by the convolutional layers are globally correlated. There are also several kernels present in each layer that is completely linked. However, when working with thick layers, these kernels will only be applied to the input map once. Due to the fact that each kernel is only used once, thick layers do not need a significant amount of processing power. The whole CNN architecture is shown here in Figure 2, which can be seen here. The input that the proposed CNN architecture receives is the pre-trained vector matrix denoted by "Y." Let “K” be the group of input vector matrix in “K” predefined vectors. The input vector matrix of each and every word is given as follows

$$Y_{1:K} = Y_1 \oplus Y_2 \oplus Y_3 \oplus Y_4 \oplus Y_k \quad Y_{1:K} = Y_1 \oplus Y_2 \oplus Y_3 \oplus Y_4 \oplus Y_k \qquad (1)$$

$\oplus$ denotes the “concatenation operation”. In the first stage, filter layers with dimension of L ϵ Y_k where “Y” is y-dimensional pretrained vectors and k is size of each filter layers and also the number of vectors in input matrix. The input word matrix is convoluted with filter kernel k to get first set of features.

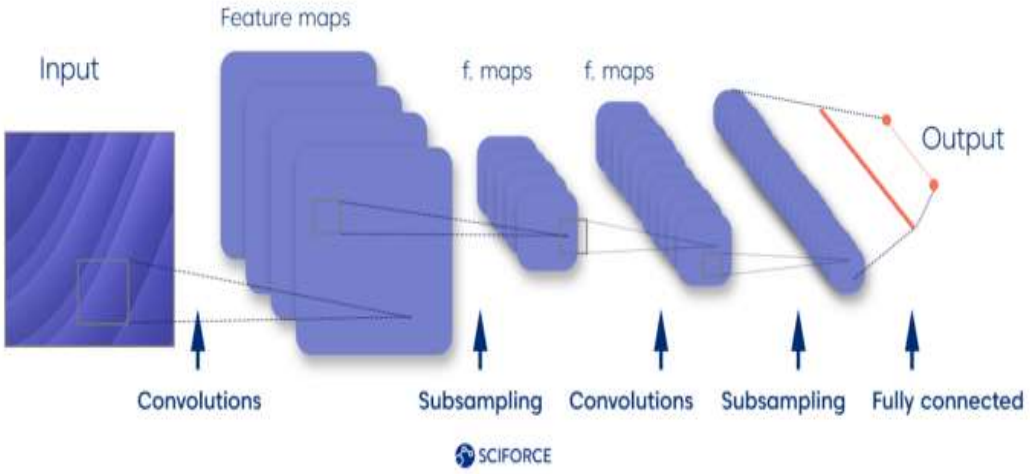

FIGURE 2. CNN Feature Extractor for Recommendation System

The output feature maps are then fed to pooling layers where pooling layers take the average value of $A(a_{avg})$, which is the most important feature with average values as the features equivalent to its particular kernel. This complete cycle is employed for one filter and to extract the feature from max- pooling layer. Multiple filter layer and pooling layers are used to attain the numerous feature maps.

TABLE 1. Most Significant Extracted Features from Social Networks Used for Training.

Centrality Features	Significance
In degree Cen.	The number of connections that are linked to the nodes are represented by this value.
Out degree Cen.	
Between Cen.	This value reflects the proportion of the shortest paths that go via each of the nodes.
Closeness Cen.	Indicates the amount of space that separates individual nodes in the networks.
Eigen Vector Cen.	Utilized in the process of calculating the centralities of other nodes found in the network
PageRank Cen.	Calculates the order of the nodes in the network based on the degree to which each node is vital to the network.
Position Cen.	Identifies the location of the nodes in relation to the nodes that are considered significant.
Clustering Co-efficient	Indicates the percentage of triangles that are present all over existing triangles in the area around the nodes in question.

The network details are 5-layers of Convolutional layers and 4-layers of Average Pooling layers along Rectified Linear Unit (ReLU) activation unit are used to extract the unknown features from database automatically. Table 1 explains the most significant Extracted Features from Social Networks used for training.

ELM Recommendation System Products Forecaster

In the hidden layer, the 'U' neurons are required to operate with an activation function that is indefinitely differentiable (like the sigmoid), but the activation function for the output layer is straight. Hidden layers in Elm should not be tuned by default.Figure.3 explains the ELM standard architecture. In the suggested ELM approach, the hidden layer does not need to be modified [20]. Different deep learning architectures are employed in [21-22] for medical image applications using image based diagnosis systems.

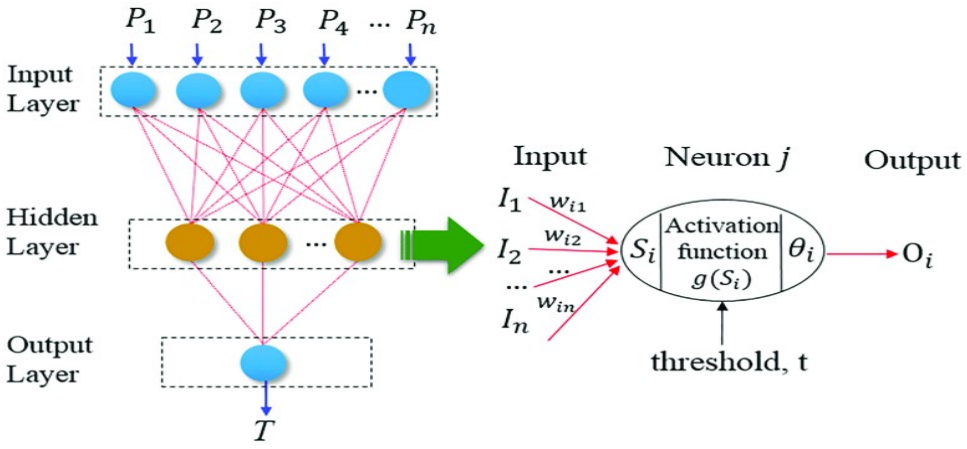

FIGURE 3. ELM Standard Architecture.

Mathematical Model of ELM

In other words, before taking care of the data from the training set. The equation that describes the system yield for an ELM with a single hidden layer may be found here (2)

$$B_K(P) = \sum_{K=1}^{L} S_K D_K(P) = C(P)S \quad B_K(P) = \sum_{K=1}^{L} S_K D_K(P) = C(P)S \text{ (2)}$$

Where P→ input

S →output weight vector and it is given as follows as

$$S = [S_1, S_2, \dots\dots\dots\dots S_L]^T (3)$$

C(P)→ output hidden layer which is given by the following equation

$$C(P) = [C_1(P), C_2(P), \dots\dots\dots\dots C_L(P) \quad (4)$$

$$C = \begin{bmatrix} C(P_1) \\ C(P_2) \\ \vdots \\ C(P_N) \end{bmatrix} \quad (4)$$

$$S' = C^* D = C^T (CC^T)^{-1} D \quad (5)$$

Where C∗→ "inverse of "C" known as Moore–Penrose generalized inverse".

$$S' = C^T (\frac{1}{N} CC^T)^{-1} D \quad (6)$$

Hence the output function can be found by using the above equation

$$B_K(P) = C(P)S = C(P)C^T (\frac{1}{N} CC^T)^{-1} D (7)$$

There is no pattern to the distribution of loads on the concealed layer (counting the bias loads). It is not accurate to say that hidden nodes are unimportant; nonetheless, it is not necessary to make any adjustments to them, and the parameters of hidden neurons may be arbitrarily selected even before the computation is performed.

RESULTS AND DISCUSSIONS

The proposed models are evaluated with the public dataset available in the Kaggle network which is utilized for both training and testing purpose. Five different labeled models such as "KNN, DT, RF, SVM, and CNN_ELM" are trained with the extracted features in order to identify the most significant or influential node in these complex networks automatically. Finally, the proposed CNN is combined with ELM algorithm to obtain more accurate results with less time complexity. All the testing is performed on an HDD that has a "Intel I7CPU with 2GB NVIDIA GeForce K+10 GPU, 16GB RAM, 3.0 GHZ," and a total capacity of 2 terabytes (TB). Tensor Flow 1.8 with the Keras application programming interface is used to build the suggested architecture. Python version 3.8 is used for the majority of the coding in each and every one of the anaconda-based apps.

Using the many datasets that were acquired, performance measures for the suggested framework such as "accuracy, precision, recall, specificity, F1-score, and specificity" were analyzed. An explanation of how to estimate performance metrics via the use of mathematical expressions is provided in Figure 4.

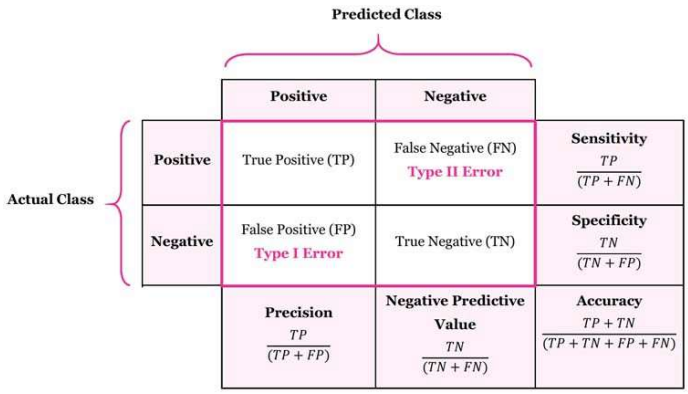

FIGURE 4. Estimation of Performance Metrics Using Mathematical Expressions.

The numerical methods for computing performance measurements are shown in the chart below. Training data is assumed as 70% and testing data as 30% from each database. The recommendation system predictor is designed to analyze the sentiments, opinions, and emoji features to detect the best upcoming products and increase the online stores' revenue. The performance parameters are detailed above which is obtained for 30% testing unseen data.

TABLE 2. Comparison Results obtained for Proposed Model for Three Datasets.

Algorithm Details	Performance Metrics				
	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1-Score (%)
WhatsApp Datasets	**96.50%**	**93.80%**	**90.20%**	**93.20%**	**94.50%**
Facebook datasets	**96.1%**	**93.20%**	**90.50%**	**92.40%**	**93.50%**
Twitter datasets	**96.4%**	**94.20%**	**91%**	**94.60%**	**93.70%**

In total, three different datasets are used such as "Amazon, Kaggle, and Twitter", for preparation and evaluation. Table 2 explains comparison results obtained for proposed model for three datasets. The below tables and figures are illustrates the obtained results.

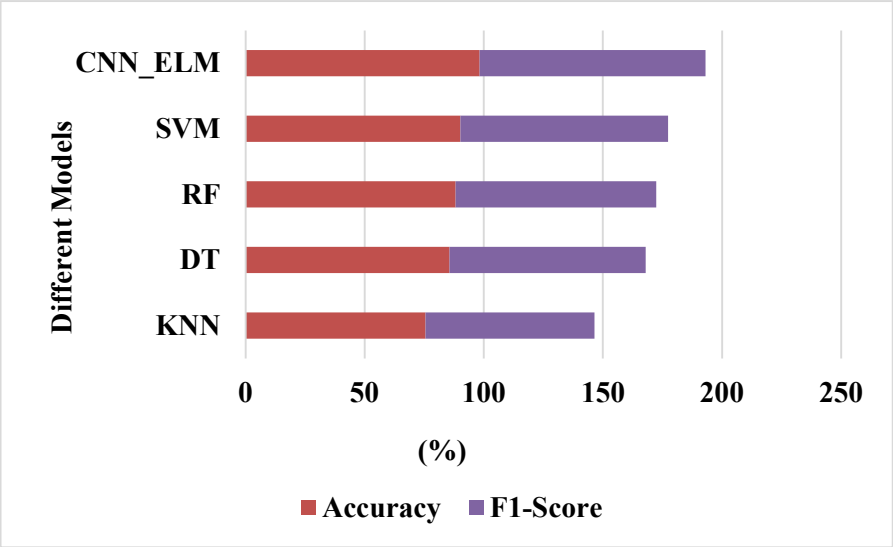

FIGURE 5. Performance Metrics for WhatsApp database.

Figure 5 explains the Performance metrics for WhatsApp database based on accuracy and F1-Score. Here, presents the percentage of accuracy and F1-Score of KNN, DT, RF, SVM, and CNN_ELM as supervised learning models for Performance metrics for WhatsApp database. From this figure, CNN_ELM model perform better than other learning algorithms.

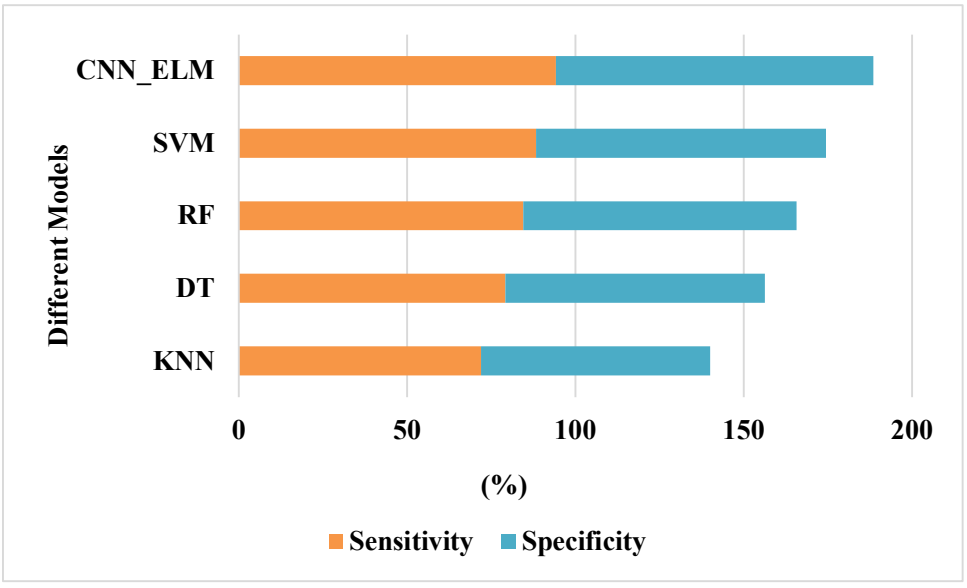

FIGURE 6. Performance metrics for WhatsApp Database.

Figure 6 explains the Performance metrics for WhatsApp database based on sensitivity and specificity. Here, presents the percentage of accuracy and F1-Score of KNN, DT, RF, SVM, and CNN_ELM as supervised learning

models for Performance metrics for WhatsApp database by sensitivity and specificity. From this figure, CNN_ELM model perform better than other learning algorithms.

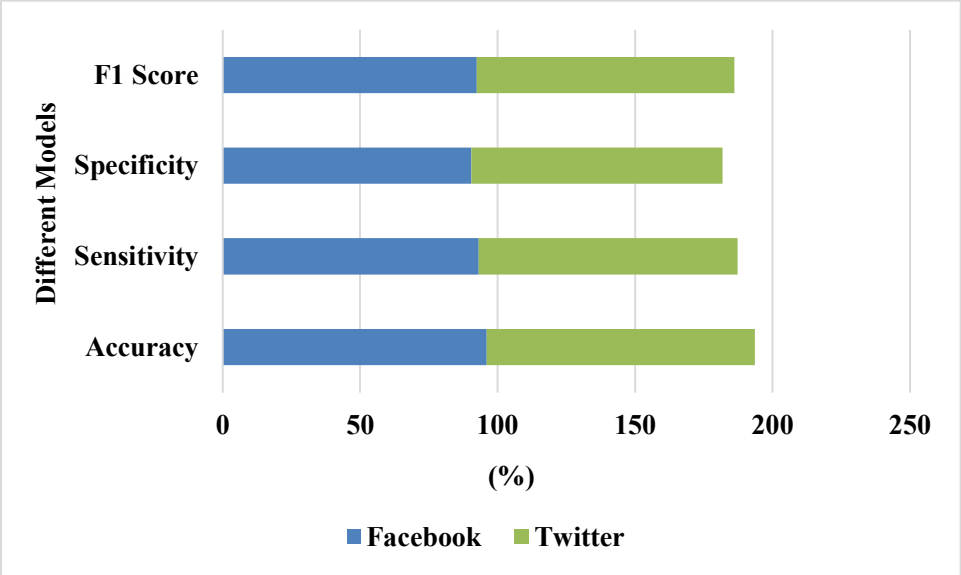

FIGURE 7. Performance Metrics for Two Distinct Database.

Figure 7 explains the Performance metrics for two distinct databases. Here, presents the percentage of Facebook and Twitter of KNN, DT, RF, SVM, and CNN_ELM as supervised learning models for Performance metrics for Facebook and Twitter. From this figure, CNN_ELM model perform better than other learning algorithms.

CONCLUSIONS

In the research field, recognizing prominent nodes in complicated systems has gotten a lot of attention. In several contexts, however, there is always a complicated and nonlinear relationship between a node's functional importance and its many properties, such as local location, worldwide location, and so on, which is difficult to characterize by a single centrality. Due to the complex structure of social networks and the large number of nodes that are interconnected with one another, the identification of prominent nodes has become an extremely important procedure. In order to overcome this obstacle, the author of this study proposes a method that is driven by machine learning and uses scenarios that include propagation to determine how significant individual vertices are. When it comes to supervised learning models, such as KNN, DT, RF, SVM, and CNN ELM, distinct supervised predictors are constructed in order to identify the nodes in complex social networks that have the greatest influence on the network as a whole. In terms of accuracy, sensitivity, and specificity, as well as the F1-Score measure, the CNN ELM model that was developed performed better than alternative learning algorithms. The average F1-score that CNN ELM achieved was 94.2%, and its detection accuracy was approximately 98.5%.

REFERENCES

[1]. M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley and HA Makse, 2011, "Identification of influential spreaders in complex networks," *Nature Physics*, 6**(11)**, pp. 888–893.

[2]. D. Chen, L. Lu¨, M. S. Shang, Y. C. Zhang and T. Zhou, 2012, "Identifying influential nodes in complex networks," *Physica A Statistical Mechanics & Its Applications*, **391(4)**, pp. 1777–1787.

[3]. X. Zhang, J. Zhu, Q. Wang and H. Zhou, 2013, "Identifying influential nodes in complex networks with community structure [J]," *Knowledge-Based Systems*, **42(2)**, pp.74–84.

[4]. B. Hou, Y. Yao and D. Liao, 2012, "Identifying all-around nodes for spreading dynamics in complex networks," *Physica A Statistical Mechanics & Its Applications*, **391(15)**, pp. 4012–4017.
[5]. P. Basaras, D. Katsaros and L. Tassiulas, 2013, "Detecting Influential Spreaders in Complex, Dynamic Networks," *Computer*, **46(4)**, pp. 24–29.
[6]. A. Zeng and C. J. Zhang, 2012, "Ranking spreaders by decomposing complex networks," *Physics Letters A*, **377(14)**, pp. 1031–1035.
[7]. J. G. Liu, Z. M. Ren and Q. Guo, 2014, "Ranking the spreading influence in complex network," *Physica A Statistical Mechanics & Its Applications*, **392(18)**, pp. 4154–4159.
[8]. L. Lu¨, Y. C. Zhang, H. Y. Chi and Z. Tao, 2011, "Leaders in Social Networks, the Delicious Case," *Plos One*, **6 (6)**, pp. 1-9
[9]. Y. B. Zhou, L. Lu and M. Li, 2011, "Quantifying the influence of scientists and their publications: Distinguish prestige from popularity," *New J. of Physics*, **14(3)**, pp. 1-17
[10]. L. Lu¨, D. B. Chen and T. Zhou, 2011, "Small world yields the most effective information spreading," *New J. of Physics*, **13(12)**, pp.825–834.
[11]. A. Shimbel, 1953, "Structural parameters of communication networks," *Bulletin of Mathematical Biology*, **15(4),** pp. 501–507.
[12]. T. Xu, J. Chen and Y. He, 2004, "Complex Network Properties of Chinese Power Grid," *Int. J. of Modern Physics B*, **18(17n19)**, pp: 2599-2603.
[13]. C. H. Wu, 2016, "A complex NLR signalling network mediates immunity to diverse plant pathogens," *Proc Natl Acad Sci U S A*, **114(30)**, pp. 8113-8118.
[14]. S. Murugan, A. Bhardwaj, and T.R. Ganeshbabu, 2015, "Object recognition based on empirical wavelet transform," *Int. J. of MC Square Sci. Res.*, **7(1)**, pp. 74-80.
[15]. T.R.G. Babu and B. Pattanaik, 2021, "Space Vector Modulation Control Based Induction Motor for Photovoltaic Application," *Int. J. of MC Square Sci. Res.*, **13(4)**, pp. 1-6.
[16]. F. Esayas, 2018, "Non-Linear Test Pattern Generation for Stuck at Fault Identification," *Int. J. of MC Square Sci. Res.*, **10(1)**, pp. 11-16.
[17]. N. Manahoran and Srinath, 2018, "M V K-Means Clustering Based Marine Image Segmentation," *Int. J. of MC Square Sci. Res.*, **10(1)**, pp. 32-37.
[18]. S. Murugan, S. Mohan Kumar and T. R. Ganesh Babu, 2020. CNN model Channel Separation for glaucoma Color Spectral Detection. *Int. J. of MC Square Sci. Res.*, **12(2)**, pp. 27-36.
[19]. K. He, X. Zhang, S. Ren S and J Sun, 2016, "Deep Residual Learning for Image Recognition," *2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society*, pp. 770-778
[20]. S Murugan, TR Ganesh Babu and C. Srinivasan, 2017, "Underwater Object Recognition Using KNN Classifier," *Int. J. of MC Square Sci. Res.,* **9(3)**, pp. 48-52.
[21]. M. A. Ramitha and N. Mohanasundaram, 2021, "Classification of Pneumonia By Modified Deeply Supervised RESNET and SENET Using Chest X-Ray Images," *Int. J. Adv. Sig. Img. Sci*, **7(1)**, pp. 30–37.
[22]. M. Alagirisamy, 2021, "Micro Statistical Descriptors for Glaucoma Diagnosis Using Neural Networks," *Int. J. Adv. Sig. Img. Sci*, **7(1)**, pp. 1–10.