Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 44 No. 5 (2023)

Statistical Authentication Technique for Facebook using Social Network Analysis

[1]Shalini Hanok, [2]Navya N, [3]Kavyashree M K, [4]Anupama S
[1][2]Dept. of ECE, ATME College of Engineering,
Mysore, India
[3][4] Dept. of ECE, SJCE, JSS S&TU,
Mysore, India

Abstract—People spend the majority of their browsing time on social networking sites (SNS) like Facebook, Twitter, WhatsApp, and others in the modern world, which increases the magnitude of sensitive personal information that is shared online and spreads quickly. One among the many threats that internet criminals bring about that must be avoided is the development of fictitious Facebook profiles. The human brain has the most powerful capacity for forming decisions based on social contact, visual signals, contextual information, and spatiotemporal data. To enhance the functionality of the current biometric system, an automated intelligent biometric system is built in this research considering the social behavioral traces found on online social networks (OSNs). These traces are dependent on human thinking ability. This technique is employed to establish unique Facebook accounts. Results from an experiment using the Facebook database showed a genuine acceptance rate (GAR) of 99% and a false acceptance rate (FAR) of 1.6% with 99% accuracy.

Index Terms—Online Social Network (OSN), Social Network Analysis (SNA), Social behavioral biometric (SBB), Social net-working site (SNS)

1. Introduction

Since it plays such a significant role in everyone's everyday lives, the internet is the focal point of the contemporary era of the twenty-first century. Many people use SNS to communicate, in addition to sharing their private and sensitive data as images, videos, buddy lists, and conversations. Popular SNS include Facebook, WhatsApp, LinkedIn, and other well-known SNS. Despite the fact that implementing SNS on OSNs has several benefits, there are some privacy and security issues that must be resolved. Identity theft, identity fraud, and impersonation are some of the key security issues that SNS users now face.

The necessity for reliable methods to identify authorized users stems from the simple reason that the attackers or stalkers frequently accessing private details from SNS accounts, breaching both the SNS's security and privacy. Numerous strategies have been devised to safeguard the user's identity. For instance, Facebook stores the standard internet protocol (IP) addresses and gadgets utilized by each account when logging into its website, and it checks to determine if an identical account has logged in on additional IP addresses or devices by posing a few confidential queries [1], sending security codes to the relevant mail, or sending a One Time Password (OTP) to the user's phone number registered with it to confirm their identity [2].

Facebook accounts come with a range of privacy settings to give the user privacy. Fake accounts are still being created on Facebook despite all the precautions taken to safeguard user safety. Facebook serves as one of many well-known OSNs worldwide. Since Facebook is so widely used, numerous illicit enterprises have emerged to access users' Facebook accounts and track data like likes, comments, and shares, among other things. These evil deeds include the creation of fraudulent accounts that launch online assaults on OSNs [2]. Several inquiries have been conducted to detect users of fraudulent accounts, but none of them have been successful in stopping the creation of fake accounts.

Facebook classifies fake accounts as duplicate accounts or fraudulent accounts. Duplicate accounts are those made by a user added to their primary accounts. Unauthorized accounts and user-misclassified accounts are the two categories of false accounts. Unwanted accounts are made to violate Facebook policies, such as spamming, whereas users establish user-misclassified accounts to develop personal profiles for businesses, organisations, etc. False accounts are purposefully made in order to manipulate or manage a specified community or a bunch of individuals in an illegal manner [3]. Thus, using automated biometric is among the distinctive strategies suggested to prevent the establishment of false accounts.

A sophisticated computer initiative that duplicates human vision, identification, and decision-making is an automated biometric authentication system for detecting a person [4]. Researchers predicted that as intelligent computer systems developed that biometric systems would be capable to accurately replicate human cognitive processes eventually [5].

In the growing discipline of social behavioural biometrics (SBB), human behaviour on OSNs generates distinct differences which may be utilised to individually identify each person. The identification in the suggested method is based on the user's Facebook activity, interactions with other users, and analysis of individual information given in the social network. Establishing distinguishing characteristics that characterise user behaviour is necessary for this. [6].

2. Literature Review

During the last few years, there has been a notable in-crease in the number of online social relationships. Daily social interactions on the Internet create a significant amount of behavioral traces that are very challenging to imitate or change. This prompts the proposed study to offer web-based SBB and their empirical investigation of person authentication. The danger to users of these websites is rising in tandem with the growth in the popularity of several online SNSs including Facebook, Twitter, LinkedIn, and Instagram. The most prominent community-building website is Facebook, according to statistical analysis, because so many people have accounts online and use it regularly. In 2020, there will be 2.70 billion active Facebook users, per the report. Therefore, if this website is not secured, its 2.70 billion visitors are at risk [7].

The security of users' personal information is significantly threatened by fake or duplicate Facebook accounts. Academics have developed a number of techniques that are effective at spotting fake Facebook profiles in order to counter this threat. The approaches that have been researched are listed below. The spam detection of comments on Facebook social network, Sohrabi et al [8] used supervised machine learning, clustering, and decision tree algorithms that examined character attributes found in comments and postings of individual accounts. The suggested strategy had a 91.2% accuracy rate.

In a real-time scenario, Dewan et al [9] proposed automatic analysis of dangerous information on Facebook using a broad range of requirements according to the individual's profile, text data, metadata, and URLs. Based on their findings, they were able to attain an accuracy of 86.9% using machine learning models to train the dataset.

Shan-Hung Wu et al. [10] state that they presented a instance of Facebook for fraud detection in their research and that they felt this study violated the privacy of other people's Facebook accounts by using the data to identify fraud. the identical hardware and IP address, the author created a mechanism to track down stalkers who had logged into the Facebook accounts of account holders. In two minutes, the results were 80% correct, and in seven minutes, 90% accurate. A 5% false alarm rate and 93.53% accuracy were achieved by the Receiver Operating Characteristic (ROC) curve during the accuracy test. Authenticity and fraud must be distinguished by using original social network profiles,

Wani et al [11] devised a technique for detecting fake Facebook profiles by examining emotive elements present in an individual's Facebook page. On publicly accessible Facebook postings from individuals, the article explored with various detection model approaches, including SVM, naive bayes, and random forest machine learning algorithms. The four detection model techniques mentioned above produced results of 87.66%, 83.44%, 85.71%, and 90.9%, respectively. Twelve emotion-based features are used to train the detection model.

Albayati et al [12] discovered fraudulent Facebook profiles using a variety of data mining technologies. The ID3 decision tree, k-NN, and SVM, unsupervised machine learning methods, including k-means and k-medoids, were employed by the researcher to obtain accuracy of 97.76%, 91.45%, and 95.72%, respectively.

Logistic regression had been used as a machine learning technique by Xiaochen Hu et al [13] to analyse the data. The Bureau of Justice Statistics (BJS) showed that overall, 91.3% of cases were correctly classified using the National Crime Victimization Survey Identity Theft Supplement (NCVS-ITS) in 2012, 2014, and 2016.

Data mining techniques to protect authentic user data and recognize untrue accounts on SNSs are highlighted by M. Senthil Raja et al [14]. The method to conduct this research which is to recognize untrue accounts on on OSN is based on the deceptive person's frequent sharing of posts throughout the day additionally their most recent behaviour and activity, as determined by the analysis related to 3PS (Publicly Privacy Protection System). Creation of OSN accounts for demonstrations, keeping informed on the most recent postings, com-ments, and photos, searching

the web, along with additional appropriate measures are covered in this activity. This method has a 94.6 % accuracy rate, it is discovered.

3. Proposed Authentication Method

The access control technique used for logging into the Facebook social network is depicted in Fig. 1. When an individual first connects to the SNS server, the Social Network Analysis (SNA) of the user's friend network features is run. The attributes are subsequently provided to a verification system, which, in the event if the individual is authorized user, the access is permitted. If any suspect or fraudulent behaviour is discovered, the user will refuse to authenticate the Facebook account.

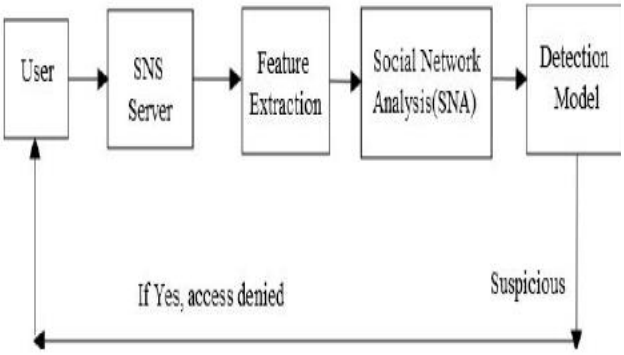

Fig. 1: Basic Flow Diagram

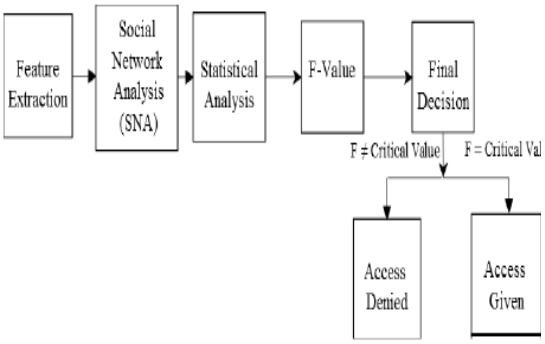

Fig. 2: Detection Model of SNS

The features gathered from the user and the features kept in the repository are statistically compared by obtaining the F-test values, as displayed in Fig. 2. The access is regarded as a valid user when the F-Value equals the specified value; contrary, they are considered predators.

4. Dataset Modelling

To construct this model, the friendship network from all the Facebook users are gathered and SNA is performed to the collected features, the features collected are stored in database as training set data, test sample and training set samples are compared to obtain F-Value using a statistically significant F-test evaluation. Whenever the F-Value is One and an individual is granted access, the testing data is acknowledged as a legitimate user; otherwise, if the F-Value is not 1, the test sample is identified as a unauthorized user and the access is denied.

A. Feature Extraction

In our case study, among the most well-known social network platforms, Facebook, is utilized to collect distinctive patterns of people so that you can find them on social networks. A specific Facebook profile can be identified using a number of distinctive characteristics, with Friendship Network (FN) being one such characteristic.

1) Friendship Network: From a person's Facebook Friend-ship Network, the FN is derived. Friends are represented by Node 2, 3, 4, 5, 6, 7, and 8 in Fig. 3, which shows the FN of Facebook user 1.

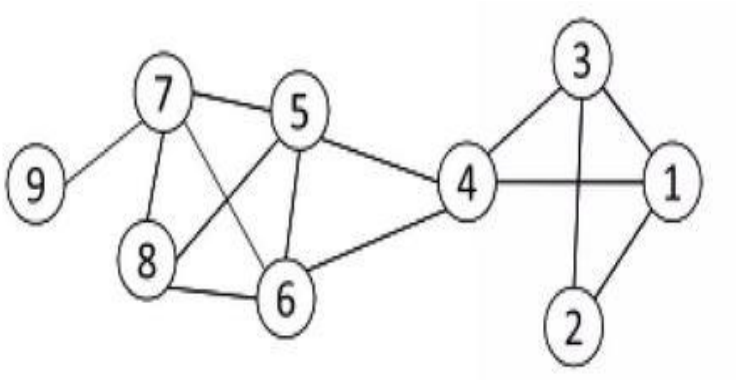

Fig. 3: Friendship Network Illustration

2) Data Collection from Facebook: Consider taking a look at the FN1 in Fig. 4 for Person 1. This FN1 is subjected to some SNA measurements, and training data is used to validate the outcomes. The data table produced from the FN in Fig. 4 is displayed in Fig. 5 and contains several computed parameters.

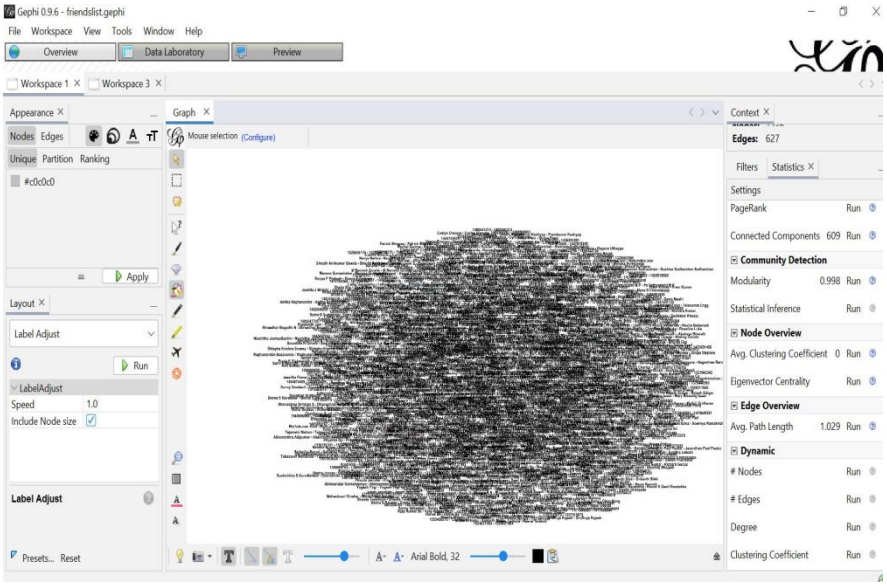

Fig. 4: Illustration of FN1 of Facebook profile P1 visualized using Gephi

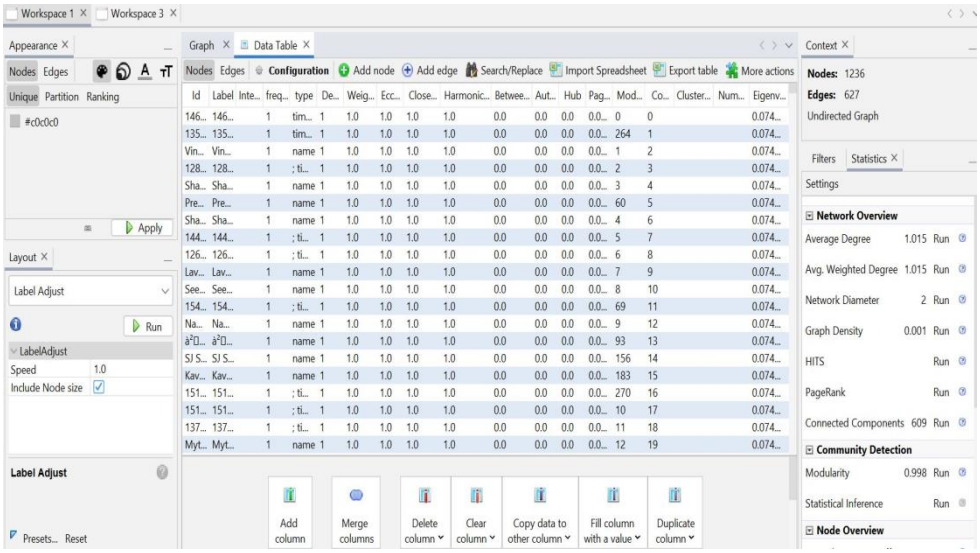

Fig. 5: Table of data from the FN1

After extracting the FN from the features from Facebook profile a couple measurements are made using FN network which are discussed below:

• **Average weighted degree:** Weights added together on the margins of nodes on average. The total amount of times a set of vertices have crossed a region in the plot is represented by its weight. If a node's weight is higher, it has likely had more visits than nodes with lower weight degrees. The weighted degree of the node is comparable to the degree. It depends on a node's amount of the edges, but it also takes into account the weight of each edge. It is exerting the weight of the edges added together.

$$A_W(i) = \frac{\sum_{g \epsilon FN(n)} W_g}{N} \qquad (1)$$

• **Network Diameter (ND):** As an additional method of measuring network graphs, the diameter of a network can be defined as the longest of all the found shortest paths. The shortest possible path exists within the network connecting the two closest nodes. The diameter is the largest of all predicted path lengths, and the diameter has users been determined to be the nearest path length between one node and all other nodes. The diameter acts as an indicator for the linear size of a network. If nodes A, B, C, and D are connected, the diameter from A to D would be three (3-hops, 3-links) [15].

• **Modularity:** The modularity of a network refers to how much this may be broken down into communities or clusters. Networks having an advanced level of modularity have strong links between nodes inside every module and sparse connections with nodes in other modules. Formally, modularity is defined as the portion of edges that adhere to one of the specified groups minus the anticipated portion of edges that would do so if edges were distributed randomly. Modularity is defined by

$$P(ij) = \frac{1}{2n} \sum_{ij} (B_{ij} - \frac{G_j G_i}{2n}) \delta_{D_i D_j} \qquad (2)$$

A network's ability to be divided into modules was intended to be measured by the concept of modularity (also called groups, clusters or communities) [15].

• **Average Cluster Coefficient:** The "all-my-friends-know-each-other" feature is quantified by the clustering coefficient. The expression used to illustrate this is "My connections are the connections of my neighbours". The ratio of actual linkages tying a node's neighbours together by the overall count of a possible links is the clustering coefficient of a node, to put it more accurately. The average of each node's clustering

coefficients makes up the network's overall clustering coefficient. A network's high clustering coefficient is another sign of a small universe [15]. The n-th node's clustering coefficient is

$$Q_n = \frac{2a_n}{m_n(m_n - 1)} \quad (3)$$

where,

m_n = the nth node's number of neighbors.

a_n= interactions among these neighbors.

Therefore, the highest number of connections between neighbors is

$$\begin{Bmatrix} m \\ 2 \end{Bmatrix} = \frac{m(m-1)}{2} \quad (4)$$

- **Average Path Length:** In each pair of two nodes, the shortest route connecting them is identified, and its length is averaged over all other paths to calculate the typical shortest path length, the length is the quantity of path-included intermediate edges i.e., The number of intermediate edges in a path defines how long it is, or by the distance $d_{u,v}$ between two graph vertices u, v. This demonstrates the average amount of steps required to go between a single network node to other.

The above parameters are measured from 20 different users and tabulated in Table I

Table 1: Calculating parameters for 20 users

Database data extracted from 20 different account							
Sl.No	Account Owners	Friendship Network	$A_W(i)$	ND	P(ij)	Q_n	$d_{u,v}$
1	P1	FN1	21.846	17	0.835	0.303	4.338
2	P2	FN2	16.218	18	0.923	0.432	5.432
3	P3	FN3	15.021	8	0.747	0.647	3.501
4	P4	FN4	1.015	2	0.998	0	1.029
5	P5	FN5	14.386	15	0.729	0.357	3.889
6	P6	FN6	2.025	3	1.234	0.135	1.324
7	P7	FN7	13.275	12	0.645	0.321	1.245
8	P8	FN8	11.432	11	0.545	0.233	4.543
9	P9	FN9	3.456	5	0.945	0.231	1.543
10	P10	FN10	1.234	2	1.003	0.123	1.641
11	P11	FN11	17.234	15	0.754	0.278	6.213
12	P12	FN12	10.765	11	0.643	0.356	4.125
13	P13	FN13	3.032	2	0.789	0.176	1.578
14	P14	FN14	5.489	3	0.654	0.156	1.544
15	P15	FN15	9.432	13	0.894	0.254	3.874
16	P16	FN16	10.293	9	0.478	0.257	4.378
17	P17	FN17	3.567	5	1.678	0	1.23
18	P18	FN18	5.789	3	2.034	2	2.678
19	P19	FN19	7.89	4	1.478	1	1.645
20	P20	FN20	9.125	1	0	3.167	1.789

B. F-Test to Calculate F-Value

Most frequently, it is employed to evaluate statistical approaches that have been fitted to data sets and decide which model more closely resembles the community that the information was gathered from. To be able to analyze the actions involving two variances from samples at random samples taken from both separate normal populations, Fisher developed the F-distribution. The null theory is either supported or refuted using the F-test. The F-Value between the test sample and training sample is calculated using equation 5.

$$A = \sum_{j=1}^{m} \frac{b_j(\overline{T_j} - \overline{T})^2}{m-1} \tag{5}$$

Where, $\overline{T}_j$ tands for the jth group's sample mean.

b_j is the amount of evaluations in the jth group.
$\overline{T}$ stands for the data's aggregate mean.
m represents the total number of groups.

The F-Values of the test samples and the training samples are calculated using Algorithm 1

Algorithm 1 Statistical Analysis of SNA

Step 1: Feature extracted from SNS

$$T_{ij}= \begin{bmatrix} T_{11} & T_{12} & T_{13} & \ldots\ldots & T_{1n} \\ T_{21} & T_{22} & T_{23} & \ldots\ldots & T_{2n} \\ T_{31} & T_{32} & T_{33} & \ldots\ldots & T_{3n} \\ . & . & . & \ldots\ldots & . \\ . & . & . & \ldots\ldots & . \\ . & . & . & \ldots\ldots & . \\ . & . & . & \ldots\ldots & . \\ T_{m1} & T_{m2} & T_{m3} & \ldots\ldots & T_{mn} \end{bmatrix}$$

Step 2: Calculated matrix using F-test

$$A_{ij}= \begin{bmatrix} A_{11} & A_{12} & A_{13} & \ldots\ldots & A_{1n} \\ A_{21} & A_{22} & A_{23} & \ldots\ldots & A_{2n} \\ A_{31} & A_{32} & A_{33} & \ldots\ldots & A_{3n} \\ . & . & . & \ldots\ldots & . \\ . & . & . & \ldots\ldots & . \\ . & . & . & \ldots\ldots & . \\ . & . & . & \ldots\ldots & . \\ A_{m1} & A_{m2} & A_{m3} & \ldots\ldots & A_{mn} \end{bmatrix}$$

if $A_{ij} \doteq 1$ **then**
$AUTHENTICUSER$
else
$STALKER$
end if

5. Simulations Using R-STUDIO

The simulations are performed using the R-Studio IDE, which employs R programming language. A CSV file contains the F-values, sometimes called expected results, that are obtained from training and test sets of data. Given the actual data, the expected values are categorized using a confusion matrix.

Whenever the output may be classified into two or more groups, the confusion matrix is a performance statistic for classification issues in machine learning. The table contains four separate sets of actual and anticipated values. Recall, precision, specificity, and accuracy are the four most crucial aspects of an AUC-ROC curve to evaluate. Fig 6 is the standard table used to classify data [16].

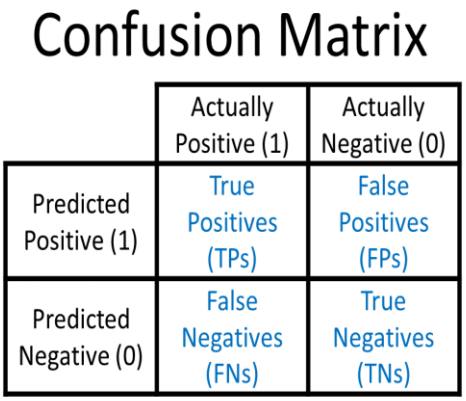

Fig. 6: Standard Confusion Matrix

The confusion matrix is explained in terms of True Positive (TP), False Positive (FP), False Negative (FN) and True Negatives (TN).

- True Positives (TPs): If the prediction came true and was positive.
- True Negatives (TNs): If the outcome is as predicted, it is true.
- False Positives (FPs): If the prediction was positive but false.
- False Negatives (FNs) : If a negative outcome was projected, it was false.

The confusion matrix and cross table confusion matrix simulation for the training and test data are depicted in Fig. 7 and Fig. 8

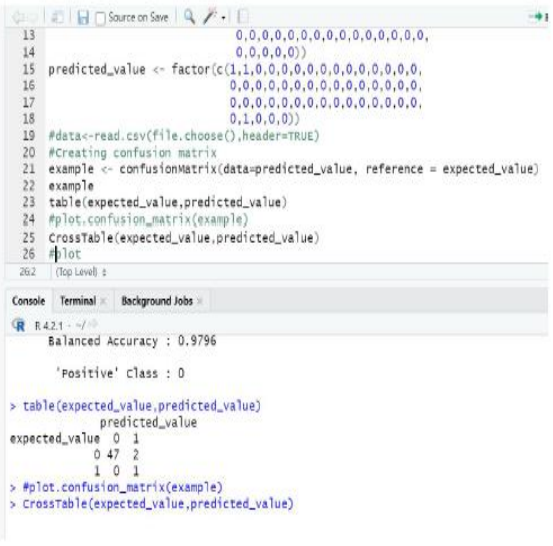

Fig. 7: Confusion matrix created from training

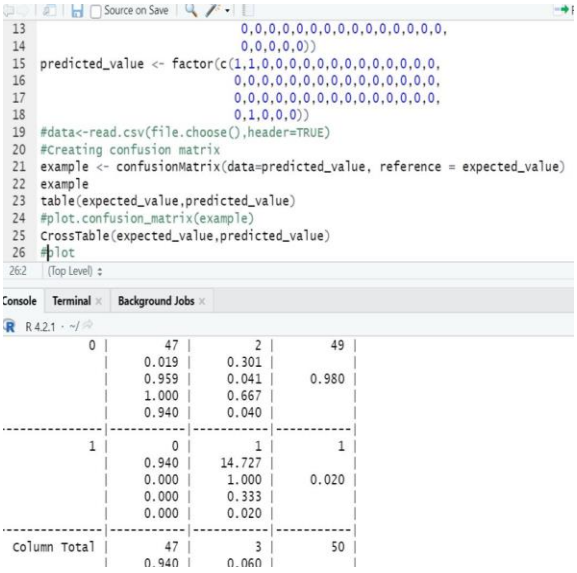

Fig. 8: Cross table of confusion matrix

6. Performance Evaluation

The effectiveness of the detection model is demonstrated via a Receiver Operating Characteristic (ROC) curve. Plotting the Genuine Accept Rate (GAR)/True Positive Rate (TPR) at the y-axis and the False Accept Rate (FAR)/False Positive Rate (FPR) at the x-axis results in the ROC curve [17].

A. False Accept Rate (FAR)/ False Positive Rate (FPR):

The FAR measures the likelihood that a biometric authentication system recognizing an incorrect access attempt made by a person who is not authorized. Divide the count of erroneous acceptances by the complete sum of identification attempts to arrive at a system's FAR [17].

$$FAR(\%) = \frac{FP}{FP + TN} \qquad (6)$$

Where, FP=Score of an impostor above the threshold.

False Positive(FP)+True Negative (TN)=Total number of attempts.

B. False Rejection Rate(FRR)

The proportion of cases that are incorrectly rejected to all tries is known as the FRR.

$$FRR(\%) = \frac{FN}{FN + TP} \qquad (7)$$

Where, FN=True scores surpass the threshold.

False Negative(FN)+True Positive(TP)= Overall quantity of tries.

C. Genuine Acceptance Rate(GAR)

The system defines the GAR as correctly acknowledged users. It is defined by

$$GAR\ (\%) = 100 - FRR\% \qquad (8)$$

Fig. 9 shows the ROC curve for the approach with a a 98% TPR and a 1.6% FPR. The graph Fig. 10 shows the TPR and TNR with respect to various thresholds, moreover, the graph demonstrates that at a threshold value of 0.5, 99% TPR, 99% TNR, and 1% FPR are achieved with 99% accuracy [17].

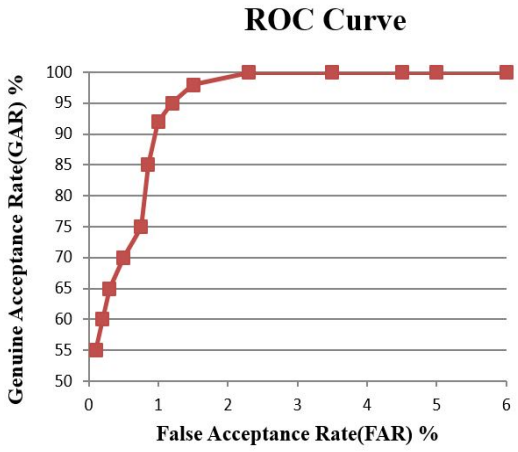

Fig. 9: Model's ROC Curve

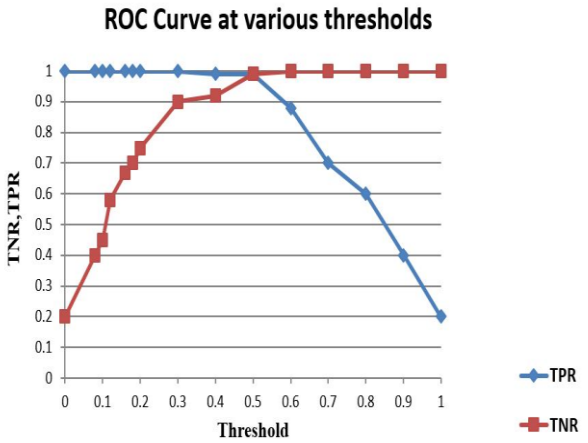

Fig. 10: ROC curve with multiple thresholds

7. Comparative Analysis

In Fig. 11 and Table II, the methodology employed in the suggested work is contrasted with the findings of several research projects conducted by researchers utilising di-verse approaches. The methodology employed in this analysis outperforms the findings including Support Vector Machine (SVM) and Karush-Kuhn-Tucker (KKT) [10], Logistic Regression (LR)[13], and the Publicly Privacy Protection System (3PS) [14].

Table II: Comparing the suggested approach to other approaches

SL.No.	Authors	Technique	Accuracy (%)
1	Shan-Hung Wu et al (2017)[10]	Support Vector Machine(SVM) & Karush-Kuhn-Tucker (KKT)	96.20
2	Xiaochen Hu et al (2020)[13]	Logistic Regression (LR)	91.30
3	M. Senthil Raja et al (2021)[14]	Publicly Privacy Protection System (3PS)	94.60
4	Proposed Method	Statistical Analysis on SNA	99

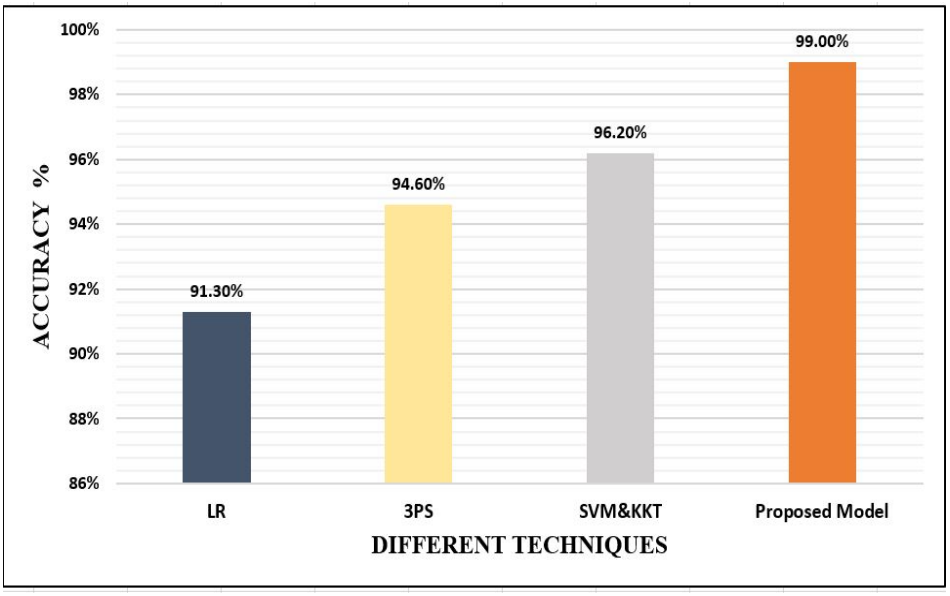

Fig. 11: Analysis of the proposed technique's accuracy in comparison to other procedures

8. Conclusion

The study suggests using a statistical analysis method on the unique attributes extracted from unique Facebook profiles to authenticate permitted individuals. When it comes to accuracy, TPR, and FPR, the suggested technique performs better than the earlier developed model. The accuracy rate achieved using a statistical analysis approach is 99%, with a TPR of 99% and an FPR of 1.6%. Other OSNs websites that request login credentials and authenticate valid users, such as Instagram, Gmail, WhatsApp, and Twitter, can put the proposed algorithm for detection developed in this research.

References

[1] Constine (2010) Facebook has users identify friends in photos to verify accounts, prevent unauthorized access online. http://www.insideFacebook.Com/2010/07/26/Facebook-photosverify/. Accessed 2012.

[2] Gupta A, Kaushal R. Towards detecting fake user accounts in Facebook. Facebook Newsroom. http://newsroom.fb.com/company-info/.

[3] Yang Z (2014) Uncovering social network sybils in the wild. Transactions on Knowledge Discovery from Data (TKDD) 8(1).

[4] Monwar MM, Gavrilova ML (2009) "Multimodal biometric system using rank-level fusion approach". IEEE Transaction System Man Cy-bernetics 39(4):867878.
[5] Gavrilova ML, Monwar M (2013) "Multimodal biometrics and intelligent image processing for security systems". IGI Glob, Hershey.
[6] https://en.wikipedia.org/wiki/Facebook [7] https://backlinko.com/Facebook-users
[8] Mohammad Karim Sohrabi, Firoozeh Karimi "A Feature Selection Approach to Detect Spam in the Facebook Social Network",Arab J Sci Eng, Springer, DOI 10.1007/s13369-017-2955, October 2017.
[9] Prateek Dewan, Ponnurangam Kumaraguru "Towards Automatic Real Time Identification of Malicious Posts on Facebook",2015 Thirteenth Annual Conference on Privacy, Security and Trust (PST), 2015.
[10] Shan-Hung Wu, Man-Ju Chou, Chun-Hsiung Tseng, Yuh-Jye Lee, and Kuan-Ta Chen "Detecting In Situ Identity Fraud on Social Network Services: A Case Study With Facebook", IEEE systems journal, vol.11, no.4, December 2017.
[11] Mudasir Ahmad Wani, Nancy Agarwal, Suraiya Jabin and Syed Zeeshan Hussain "Analyzing Real and Fake users in Facebook Network based on Emotions", 11th International Conference Communication Systems and Networks, 2019.
[12] Mohammed Basil Albayati and Ahmad Mousa Altamimi "Identify-ing Fake Facebook Profiles Using Data Mining Techniques", J. ICT Res. Appl, Vol. 13, No. 2, 2019, 107117, SSN: 2337-5787, DOI: 10.5614/itbj.ict.res.appl.2019.13.2.2, August 20th,2019.
[13] Xiaochen Hu , Xudong Zhang & Nicholas P. Lovrich " Forecasting Identity Theft Victims: Analyzing Characteristics and Preventive Actions through Machine Learning Approaches", Victims & Offenders, An International Journal of Evidence-based Research, Policy, and Practice DOI: 10.1080/15564886.2020.1806161, August 2020.
[14] Senthil Raja M, L. Arun Raj "Detection of Malicious Profiles and Protecting Users in Online Social Networks" Wireless Personal Communications, https://doi.org/10.1007/s11277-021-08095-x, January 2021.
[15] https://en.wikipedia.org/wiki/Network_science
[16] https://towardsdatascience.com/understandingconfusionmatrix-a9ad42dcfd62
[17] Shalini P, Shankaraiah "Multimodal biometric decision fusion security technique to evade immoral social networking sites for minors" Applied Intelligence,