Unmasking Fake Identities on Instagram Using Machine Learning

Olaniyi Abiodun Ayeni, Ethel Femi-Gbenoba
Department of Cyber Security, School of Computing
Federal University of Technology, Akure
Department of Information Systems and Technology
Kings University, Odeomu, Osun State
Nigeria

Abstract

The rise of fake digital identities on social media erodes trust, skews engagement, and undermines online safety. These fake accounts will help to inflate follower counts, impersonate real users, deceive advertisers, and spread scams that rake in billions of dollars from across the globe each year. Identification of such accounts is difficult because attackers often adjust their approach using synthetic images, automated posting, and fake engagement, all tailored to bypass conventional detection mechanisms. This paper aims to evaluate and compare Support Vector Machine (SVM), Random Forest (RF), and Multi-Layer Perceptron (MLP) for detecting fake Instagram accounts using profile-level data. The dataset underwent preprocessing through normalization, removal of irrelevant features, and handling of missing values. In order to address class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was applied. Multiple feature selection strategies were also employed. Recursive Feature Elimination (RFE) for SVM, feature importance ranking for RF, and mutual information scores for MLP. The dataset was partitioned using two train, validation, test splits of 70:15:15 and 80:10:10. The models were evaluated using Accuracy, Precision, Recall, F1-score. The results showed that the Random Forest classifier consistently outperformed the other models by achieving the highest accuracy and an F1-score of 0.98. SVM also delivered strong generalization performance, while MLP slightly lower in predictive strength. These findings highlight the strength of ensemble-based approaches in handling complex, high-dimensional social media data. The study contributes to ongoing efforts in developing of scalable, real-time detection frameworks to safeguard online communities as well as to preserve platform integrity.

Keywords: Fake digital identity, Instagram, machine learning, Random Forest, SVM, MLP, feature selection.

1. Introduction

The widespread use of Instagram has led to a surge in fake digital identities, threatening user trust, privacy, and platform integrity. Fake profiles are used for fraudulent activities, including scams, phishing, and fraudulent influencer endorsements [1]. These accounts will often manipulate profile features and engagement metrics, making detection challenging. Machine learning offers scalable solutions for distinguishing between real and fake accounts by analyzing user metadata and behavioral patterns.

2. Review of related works

Several studies have explored fake profile detection on social media. Dharmaraj R. Patil et al. [2]: The reason for this study was because single classifiers, like SVM or Random Forest, might fail to correctly identify advanced fake profiles. The study suggested using an ensemble model based on Majority Voting to improve detection accuracy by combining Decision Trees, XGBoost, and Random Forest. The process used a dataset of 3,474 real and 3,351 fake accounts, with features taken from activity logs and profile data. The ensemble model was tested against individual classifiers, resulting in 99. 12% accuracy. The main contribution was showing that using multiple models together made detection more reliable than using just one. However, this method used a lot of computing power and worked only with structured data, not images [2]. Pravin Kumar and Neelam [3]: The study was motivated by how fake profiles spread false information during important events like elections and pandemics. The goal was to build a real-time machine learning system that used NLP, network analysis, and user feedback. Data was collected from Twitter and Facebook via APIs, then used supervised models like Random Forest and SVM, along with unsupervised methods for detecting anomalies. The system also allowed users to report suspicious profiles. The main contribution was combining machine learning with user input to improve detection. Limitations included dependence on platform APIs which can limit data access, and the risk of many false alarms in changing environments [3]. Weronika Dracewicz and Mariusz Sepczuk [4]: The study was motivated by the fact that most current detection methods use only text features, missing out on visual clues that could help. A

 | DOI: 10.20533/ijicr.2042.4655.2025.0167

Convolutional Neural Network was developed (CNN) to classify image-based profile data. The methodology involved creating a dataset of 15,000 fake and real profile screenshots, training the CNN, and using it as a Chrome extension for real-time detection. The model achieved 96. 5% accuracy, better than Random Forest and Naive Bayes. The main contribution was the first use of images for fake account detection, showing CNNs work well for visual tasks. However, it needed manual alignment of screenshots and hasn't been tested on deepfakes [4]. Daniyal Amankeldin et al. [5]: The motivation was the increasing threat of fake accounts on social media. The goal was to create a DNN approach that achieves high accuracy with little data. A deep CNN was trained using 16 profile-based features and evaluated it with accuracy, precision, recall, F-score, and AUC. The study showed 99. 4% accuracy, which is better than traditional ML methods by around 14%. A key limitation is that it relied on a small number of features, which might not catch advanced fake profiles that mimic real users [5]. K. Harish et al. [6]: The motivation behind this study was the growing number of fake profiles on social media leading to identity theft and misinformation. The study aimed to build a model that can tell fake profiles from real ones using ML. The process included preprocessing Twitter profiles and training models like Random Forest, XGBoost, and Long Short Term Memory. The study showed that XGBoost reached 99.6% accuracy, better than other methods. A limitation was that the model had not been tested for real-time use [6]. Gaurav Vijay Barde and Dr. Nilesh R. Wankhade [7]: The motivation was the increasing number of fake accounts on platforms like Facebook, Twitter, and Instagram, often used for harmful activities. The study emphasized the need for automatic detection methods to improve platform security and user trust. The main aim was to develop an ML approach for detecting fake accounts. The process involved gathering data on real and fake profiles, extracting features from profile data and activity patterns. They tested several ML algorithms, including SVM, K-Nearest Neighbors, Random Forest, Logistic Regression, and Artificial Neural Networks, using accuracy, precision, recall, and F1-score. The main contribution was showing Random Forest achieved the highest accuracy (97.9%) for fake account detection. However, the study had limitations, including reliance on labeled datasets and not testing in real-time scenarios [7]. Shamim Ahmad and Dr. Manish Madhava Tripathi [8]: The motivation was the rise of fake profiles for cybercrimes, needing a thorough evaluation of ML techniques. The goal was to compare the effectiveness of various ML algorithms. A comprehensive review of 17 papers from 2018 to 2023 was conducted, looking at accuracy, precision, and recall. The findings showed Random Forest consistently reached over 95% accuracy across studies, with Natural Language Processing and network analysis being important for strong detection. The review added value by combining existing knowledge and highlighting the benefits of Random Forest and hybrid methods. However, it lacked original experimental validation and doesn't cover recent deep learning advances [8].

3. Motivation / problem statement

Instagram is one of the most widely used social media platforms with millions of daily active users. However, it faces significant challenges due to the rise of fake digital identities. Instagram focuses more on visual content and influencer culture. Fake accounts on Instagram are often created to inflate follower counts, manipulate engagement metrics, or impersonate real users for phishing and advertising fraud. Estimates suggest that up to 8% of Instagram accounts are fake. This contributes to billions of dollars in losses for advertisers due to fake influencer endorsements [9]. Despite advancements in detection methods from basic rule-based systems to machine learning models, these approaches face limitations. Manual detection is impractical given the scale of users, and rule-based systems require constant updating to keep up with evolving tactics used by fake profiles [9].

4. Overview of Instagram

Instagram is a social network that focuses on sharing photos and videos. It has a strong culture of influencers, where users interact through likes, comments, stories, and reels. Instagram helps people connect with friends and also lets businesses reach customers through ads and by working with influencers [10].

4.1. Key features

1. Feed posts: You can share single photos, vertical photos, carousels, or videos. You can also rearrange your profile grid and change thumbnail covers [11].

2. Instagram reels: These are short, fun videos that are easy to discover. Editing Reels has been made easier by AI, and you can try new Reels without telling your followers [11].

3. Stories: These are posts that only last 24 hours. They can also be 60 seconds long and include captions and public comments. AI tools can help change or improve images in Stories [12].

4. Direct messaging (DMs): You can chat privately with other users. Features like scheduled messages, group chats, pinned messages, translation, music sharing, and special stickers make these messages better [12].

5. Highlights: Stories can be saved in a Highlights section so they stay on your profile [12].

6. Notes and Profile Music: You can add short notes to your profile or include a music clip for visitors [12].

7. AI creative tools: The "Edits" app and other AI tools let creators make videos, add animations, change content, and convert formats quickly [13].

8. Shopping: Instagram lets users shop directly through the app and check out in-app, helping businesses and influencers reach more people [13].

9. Algorithm: It prefers original content and real interactions, like saves and shares, over just likes. When you repost something, it gets less visibility [13].

4.2. User Base and Impact

Monthly active users: 2.35+ billion in 2025 [13]. Business and marketing: Instagram is still a main platform for influencer campaigns, digital marketing, shopping, and brand visibility across many industries [13].

4.3. Challenges

Instagram has problems with fake profiles that are made to raise follower numbers, trick people, or do phishing. Fake influencer endorsements cost advertisers billions each year because bots or fake followers create fake engagement. Instagram is struggling to balance letting people express themselves with taking down bad content, and it also has to respect user privacy. The use of deepfakes makes it even harder to tell real from fake [14].

5. Model analysis

This Chapter highlights the three machine learning algorithms that will be used to detect fake digital identities on Instagram.

5.1. Support vector machines

Support Vector Machine (SVM) is a supervised machine learning algorithm that can handle classification, regression, and outlier detection problems. The idea behind the Support Vector Machines (SVM) algorithm is to find the hyperplane that best separates two classes by maximizing the margin between them. The equation for the linear hyperplane can be written as:

$$wTx + b = 0 \quad \ldots\ldots.. \quad (1)$$

where:

w = weight vector
x = input feature vector b = bias term

5.2. Random forest

A Random Forest is a collection of decision trees that work together to make predictions. It is an ensemble learning method for classification, and regression. Random forest algorithms have three main hyper parameters, which need to be set before training. These include node size, the number of trees, and the number of features sampled. From there, the random forest classifier can be used to solve for regression or classification problems.

5.3. Multi-layer perceptron

Multi-Layer Perceptron (MLP) is an artificial neural network that is widely used for solving classification and regression tasks. The network is typically organized into layers, starting with the input layer, where data is introduced. Followed by hidden layers where computations are performed and finally, the output layer where predictions or decisions are made.

Forward Propagation: In forward propagation, the data flows from the input layer to the output layer, passing through any hidden layers. Each neuron in the hidden layers processes the input as follows:

Weighted Sum: The neuron computes the weighted sum of the inputs:

$$z = \sum_i w_i x_i + b \qquad (2)$$

Where:
x_i is the input feature.
w_i Is the corresponding weight and
b Is the bias term

Activation function: The weighted sum z is passed through an activation function to introduce non-linearity.

Loss function: Once the network generates an output, the next step is to calculate the loss using a loss function. In supervised learning, this compares the predicted output to the actual label. For a classification problem, the commonly used binary cross-entropy loss function is:

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i)] \qquad (3)$$

Where:

y_i is the actual label.
$\hat{y}_i$ is the predicted label.
N is the number of samples

6. Dataset

The InstaFake Dataset consists of two subsets:

1. Fake account detection, which has a total of 1,194 accounts, with 200 being fake and 994 being real.

2. Automated account detection, which uses a richer metadata set per post.

Under Fake Account Detections, the files were further divided into 2, fakeAccountData and realAccountData. fakeAccountData contained all the fake accounts and realAccountData contained all the real accounts. These files were later combined for ease of use.

userFollowerCount	userFollowingCount	userBiographyLength	userMediaCount	userHasProfilPic	userIsPrivate	usernameDigitCount	usernameLength	isFake
25	1937	0	0	1	1	0	10	1
324	4122	0	0	1	0	4	15	1
15	399	0	0	0	0	3	12	1
14	107	0	1	1	0	1	10	1
264	4651	0	0	1	0	0	14	1

Figure 1. InstaFake fakeAccountData Sample

userFollowerCount	userFollowingCount	userBiographyLength	userMediaCount	userHasProfilPic	userIsPrivate	usernameDigitCount	usernameLength	isFake
258	238	0	0	1	0	0	10	0
263	482	30	29	1	1	0	8	0
51	78	9	0	1	1	0	10	0
297	480	22	25	1	1	2	9	0
113	242	0	95	1	1	0	10	0

Figure 2. InstaFake realAccountData Sample

Table 1. Features for Instagram dataset

Feature	**Type**	**Description**	**Fake Account Indicator**
userHasProfilPic	Binary	Whether the user has a profile picture (1=Yes, 0=No)	Fake accounts often lack profile pics or use defaults.
userIsPrivate	Binary	Whether the account is private (1=Private, 0=Public)	Fake accounts tend to avoid scrutiny by staying private.
usernameDigitCount	Numerical	Number of digits in the username	High digit counts suggest bot usernames.
usernameLength	Numerical	Length of the username (characters)	Extremely short or extremely long usernames are suspicious.
userFollowerCount	Numerical	Total followers	Fake accounts often have low followers.
userFollowingCount	Numerical	Total accounts followed	Fake accounts follow excessively or very few.
userMediaCount	Numerical	Number of posts/uploads	Low media counts suggest inactivity or spam.
userBiographyLength	Numerical	Number of characters in the bio	Fake accounts often have empty/brief bios.
is_fake	Categorical (0/1)	Indicates whether an account is fake (1) or real (0)	Primary label used to classify accounts as fake or real

* The fake account subset includes 9 features, including the label.

7. Methodology

The dataset was from InstaFake dataset [10] on GitHub. It contained 1,194 Instagram accounts, with 200 being fake and 994 being real. The dataset was then preprocessed, which involved handling missing data, removing irrelevant features, and feature scaling. Synthetic Minority Oversampling Technique (SMOTE) was applied to address class imbalance. Multiple features were engineered from the original features and were selected using Recursive Feature Elimination for SVM, Feature importance ranking for RF, and Mutual Information for MLP. The dataset was split 80:10:10 in order of training, validation, and test, respectively.

Hyperparameter tuning involved making use of Grid Search for SVM, and Random Search for RF and MLP. The models were evaluated using metrics Accuracy, Precision, Recall, and F1-score.

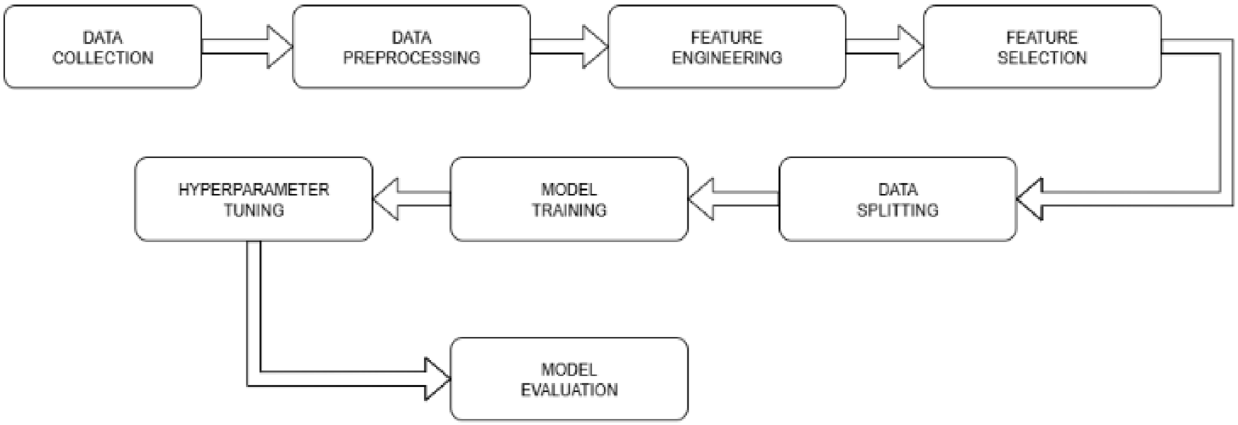

Figure 3. Model workflow

Performance Metrics

The following performance metrics were used:

Accuracy = (TP + TN) / (TP + TN + FP + FN) (4)

Precision = TP / (TP + FP) (5)

Recall = TP / (TP + FN) (6)

F1-score = 2 * ((Precision * Recall) / (Precision + Recall)) (7)

Where:
TP is true positive
TN is true negative
FP is false positive
FN is false negative.

Feature Engineering

The following features were engineered from the already existing features in the InstaFake Dataset.

Follower-Following Graph Features:

1. followers_to_following_ratio
2. following_to_followers_ratio
3. social_engagement_score

Authenticity Features:

1. has_bio
2. bio_length_to_media_ratio
3. empty_profile_score

Username Analysis Features:

1. username_digit_ratio
2. username_entropy
3. has_suspicious_username

Activity and Media Features:

1. media_to_followers_ratio
2. private_with_high_media

Results

SVM and RF show near-perfect precision (0.99) and high F1 (0.98), while MLP lags slightly (F1: 0.95) (see Figure 4).

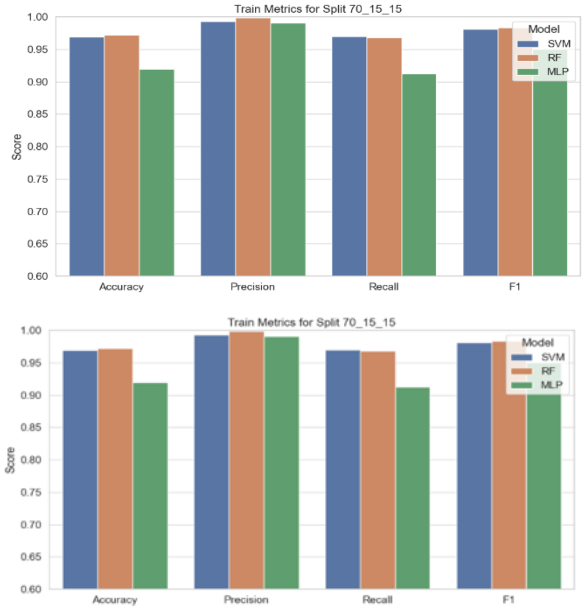

Figure 4. Model Comparison of Train Metrics for Split 70:15:15

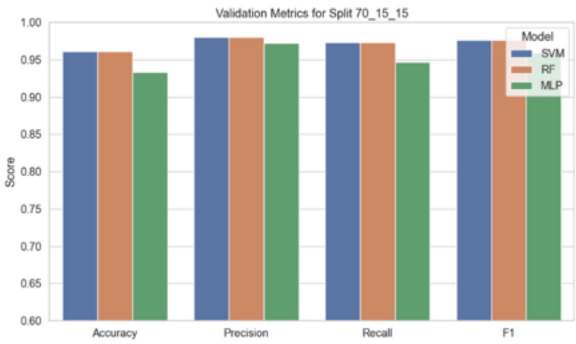

Figure 5. Model Comparison on Validation Metrics for Split 70:15:15

SVM and RF tie for top performance (F1: 0.976), with MLP close behind (F1: 0.959). All models exceed 93% recall (see Figure 5).

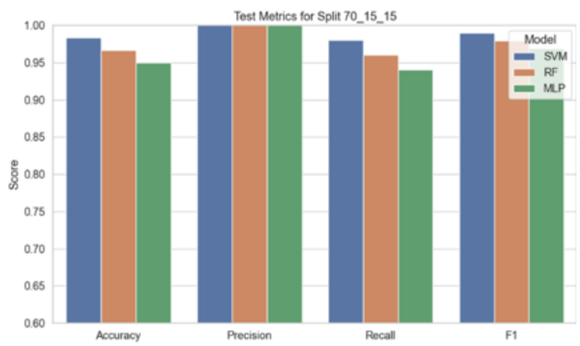

Figure 6. Model Comparison on Test Metrics for Split 70:15:15

SVM dominates (F1: 0.99, 100% precision), followed by RF (F1: 0.98). MLP maintains strong recall (0.94) (see Figure 6).

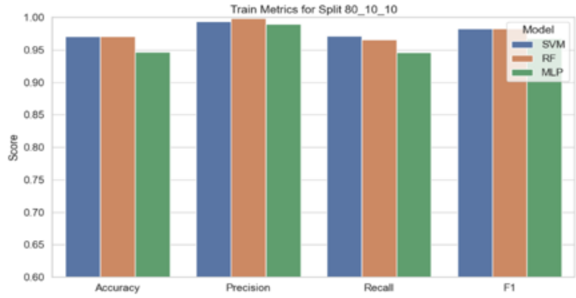

Figure 7. Model Comparison on Train Metrics for Split 80:10:10

RF edges out SVM in precision (0.999 vs. 0.994), but both share identical F1 (0.982). MLP trails (F1: 0.967) (see Figure 7).

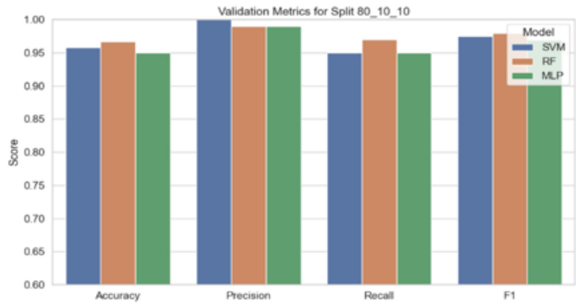

Figure 8. Model Comparison on Validation Metrics for Split 80:10:10

RF leads (F1: 0.98), while SVM and MLP tie (F1: 0.97). SVM achieves 100% precision but lower recall (see Figure 8).

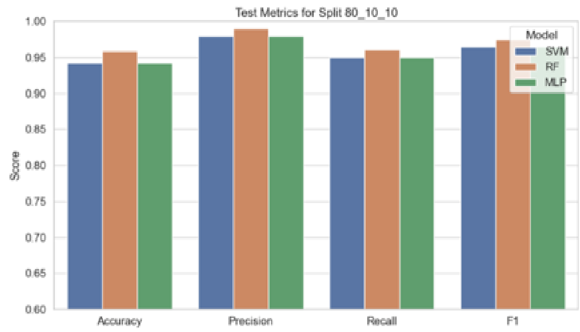

Figure 9. Model Comparison on Test Metrics for Split 80:10:10

RF performs best (F1: 0.975), with SVM and MLP closely matched (F1: 0.964). All models exceed 95% recall (see Figure 9).

Table 2. F1-score of SVM model performance

Split	Model	Train F1	Validation F1	Test F1
70:15:15	SVM	0.9811	0.9764	0.9899
80:10:10	SVM	0.9822	0.9741	0.9645

Table 3. F1-score of rf model performance

Split	Model	Train F1	Validation F1	Test F1
70:15:15	RF	0.9832	0.9764	0.9796
80:10:10	RF	0.9821	0.9796	0.9746

Table 4. F1-score of MLP model performance

Split	Model	Train F1	Validation F1	Test F1
70:15:15	MLP	0.9498	0.9592	0.9691
80:10:10	MLP	0.9672	0.9691	0.9645

Table 5. Model performance across all models

Model	Accuracy	Precision	Recall	F1-Score
SVM	0.942	0.979	0.950	0.964
RF	0.958	0.990	0.960	0.975
MLP	0.942	0.979	0.950	0.964

Random Forest achieved the highest F1-Score of 0.975. SVM produced strong generalization while MLP was slightly lower but competitive F1scores (see Tables 1 to 4).

8. Discussion

Random Forest outperformed other models due to its ability to handle complex feature interactions and class imbalance. SVM benefited from feature selection, while MLP showed efficient recall, critical for minimizing false negatives.

9. Conclusion

Random Forest demonstrated high accuracy in detecting fake Instagram profiles. Feature selection and data balancing were also key to optimal performance. Future work should integrate behavioral analysis and real-time deployment for enhanced robustness.

10. References

[1] Chakraborty, P., Shazan, M. M., Nahid, M., Ahmed, M. K., and Talukder, P. C. (2022). Fake profile detection using machine learning techniques. Journal of Computer and Communications, 10(10), Article 10.

[2] Patil, D. R., Pattewar, T. M., Punjabi, V. D., and Pardeshi, S. M. (2024). Detecting fake social media profiles using the majority voting approach. EAI Endorsed Transactions on Scalable Information Systems, 11(3).

[3] Kumar, P., and Neelam. (2024). Advanced real-time fake social media profiles detection and reporting system. International Journal of Education and Science Research Review, 11(3).

[4] Dracewicz, W., and Sepczuk, M. (2024). Detecting fake accounts on social media portals—The X Portal case study. Electronics, 13(13), 2542.

[5] Amankeldin, D., Kurmangaziyeva, L., Mailybayeva, A., Glazyrina, N., Zhumadillayeva, A., and Karasheva, N. (2023). Deep Neural Network for Detecting Fake Profiles in Social Networks. Computer Systems Science and Engineering, 47(1), 1091–1108.

[6] K. Harish, R. Naveen Kumar, and Dr. J. Briso Becky Bell. (2023). Fake Profile Detection Using Machine Learning. International Journal of Scientific Research in Science, Engineering and Technology, 10(2), 719–725.

[7] Gaurav Vijay Barde and Nilesh R. Wankhade. (2023). Social Media Fake Account Identification Using Machine Learning. International Journal of Advanced Research in Science, Communication and Technology, 261–268.

[8] Ahmad, S., and Tripathi, M. M. (2023). A Review Article on Detection of Fake Profile on Social-Media. International Journal of Innovative Research in Computer Science and Technology, 11(2), 44–49.

[9] Binder, M. (2024, February 16). The majority of traffic from Elon Musk's X may have been fake during the Super Bowl, report suggests. Mashable. https://mashable.com /article/x-twitter-elon-musk-bots-fake-traffic (Access Date: 16 May 2025).

[10] Hanna, K. T., Wright, G., and Yasar, K. (2025, January 22). What is social networking and how does it work? | Definition from TechTarget. https://www.techtarget.com/ whatis/definition/social-networking (Access Date: 19 May 2025).

[11] Richard. (2025, January 23). Learn More About Instagram's Game-Changing Features 2025. Avocado Social. https://www.avocadosocial.com/instagrams-game-changing-features-for-2025-what-you-need-to-know/ (Access Date: 21 May 2025).

[12] Pokrop, J. (2025, July 11). What's New on Instagram in 2025: New Features and Updates. NapoleonCat. https://napoleoncat.com/blog/instagram-new-features-and-updates/ (Access Date: 24 May 2025).

 | DOI: 10.20533/ijicr.2042.4655.2025.0167

[13] Social Ninjas. (2025, February 11). What's New on Instagram in 2025: A Social Media Manager's Guide [UPDATED MONTHLY]. UpGrow. https://www.upgrow.co m/blog/instagram-new-features-and-updates-2025 (Access Date: 26 May 2025).

[14] Paljug, K. (2025, February 19). Social Media: Definition, Importance, Top Websites, and Apps. Investopedia. https:/ /w ww.investopedia.com/terms/s/social-media.asp (Access Date: 26 May 2025).

[15] Akyon, F. (2022). InstaFake: Fake Instagram profile dataset [Data set]. GitHub. https://github.com/fcakyon/ instafake-dataset (Access Date: 27 May 2025).