International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

Fake Social Media Profile Detection and Reporting

Sasigaran S , Marimuthu R
Department of Cyber Forensics & Information Security
Dr MGR Educational and research institute , Tamil Nadu ,Chennai , India

Abstract **- The growing presence of deceptive accounts on social networking platforms has become a serious concern, affecting user safety, trust, and the overall credibility of these platforms. This study presents a machine learning-driven framework designed to identify fraudulent social media profiles by drawing on a diverse set of features gathered from user-generated content, behavioural indicators, and social network structures. A carefully labelled dataset containing both authentic and fake profiles is used to train and validate the system. Beyond detection alone, the framework also includes a community reporting channel that enables users to flag accounts that appear suspicious, actively contributing to the removal of harmful entities. System performance is measured through standard evaluation criteria including accuracy, precision, recall, and F1-score. The model is further verified through deployment in a real-world social media setting, confirming its practical value in limiting the spread of fraudulent accounts and improving the overall user experience.**

Keywords - fake account detection; profile verification; fraudulent profiles; reporting mechanisms; classification metrics.

I. INTRODUCTION

Social media has transformed the way people communicate, share experiences, and stay connected with the world around them. Platforms such as Facebook, Instagram, and Twitter have become central to daily life, enabling users to build communities, follow news, and interact with brands and public figures. Yet alongside these benefits, a troubling pattern has emerged: the deliberate creation of false identities designed to mislead, manipulate, or exploit other users.

A fake social media profile is essentially an account constructed to impersonate someone or to hide its true purpose. Such accounts are frequently used by scammers seeking financial gain, by individuals engaged in harassment, or by organised groups attempting to spread misinformation at scale. The consequences range from reputational damage to real financial loss, and in some cases they contribute to broader societal harm. Despite the severity of the issue, many ordinary users lack the knowledge or tools to identify and report these accounts effectively.

This research addresses that gap by developing an automated detection and reporting system. The proposed tool examines social media profiles for tell-tale signs of inauthenticity, including incomplete personal information, abnormal activity patterns, and the use of duplicated images. When a profile is flagged as suspicious, the system promptly alerts the user and provides a straightforward reporting pathway, enabling faster account removal and a safer experience for the wider community.

Instagram is given particular attention in this work due to its scale and its distinctive vulnerabilities. The platform's emphasis on visual content and follower counts makes it especially susceptible to manipulation: fake accounts frequently employ stolen photographs, artificially purchased followers, and unsolicited direct messages to deceive genuine users. By focusing on these platform-specific signals, the system aims to offer more reliable detection outcomes.

In a broader sense, this initiative moves toward making digital spaces more honest and accountable. It not only equips users with a practical detection tool but also encourages a more alert and critically aware online culture, where suspicious behaviour is neither ignored nor normalised.

II. LITERATURE REVIEW

A growing body of research has engaged with the problem of detecting inauthentic profiles across major social media platforms. The studies reviewed here collectively map the landscape of existing approaches, from classical rule-based methods through to advanced ensemble and graph-based techniques.

A. Machine Learning Approaches

Kerrysa and Utami [1] conducted a systematic survey of machine learning algorithms applied to fake account identification on Twitter, Instagram, and Facebook. Their analysis underlines the difficulties created by the open architecture of social media platforms, which allows harmful actors to proliferate fake accounts designed for spamming and misinformation campaigns. The survey concludes that no single algorithm consistently outperforms others across all platforms, pointing to the need for adaptive, context-aware solutions.

Building on this, Singh et al. [6] explored the effectiveness of Support Vector Machines and Random Forests when applied to Instagram profile data. Their findings indicate that ensemble approaches generally yield more stable results than single-classifier systems, particularly when dealing with imbalanced datasets where fake profiles are a minority class.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

B. Blockchain and Data Integrity

Deshmukh et al. [2] proposed an innovative hybrid architecture that combines machine learning with blockchain technology for the dual purpose of detection and reporting. By logging detected incidents on an immutable ledger, their system ensures that reports cannot be tampered with and that patterns of abuse remain visible over time. While the technical complexity of blockchain integration is acknowledged as a limitation for rapid deployment, the approach offers a compelling model for transparent, long-term abuse tracking.

C. Identity Deception and Graph-Based Detection

Alharbi and colleagues [3] provided a comprehensive taxonomy of identity deception attacks, distinguishing between three primary forms: the creation of entirely fabricated profiles, outright identity theft, and the cloning of an existing account to confuse followers. Their work highlights that different attack types require different detection strategies, and that most existing tools are optimised for only one or two of these scenarios.

Patil et al. [4] introduced a majority-voting ensemble that combines Decision Trees, XGBoost, and Random Forest classifiers to analyse both behavioural and profile-level attributes. The ensemble mechanism reduces the risk of systematic bias that can arise when any single model is relied upon exclusively, producing accuracy rates that surpass individual classifiers in controlled experiments.

A structurally different perspective is offered by A. K. Singh and colleagues [5], who examined social network graphs to identify fake profiles through their connection patterns rather than their content alone. Their research demonstrates that fake accounts frequently cluster together and maintain unusually sparse or unusually dense connection structures, both of which deviate from the organic patterns associated with genuine users. This graph-based signal can complement content-based features, offering a richer foundation for detection.

III. METHODOLOGY

A. System Architecture

The proposed system follows a sequential pipeline beginning with data ingestion and ending with user notification and reporting. At its core, the architecture is composed of six functional modules: data acquisition, preprocessing, feature extraction, classification, alert generation, and continuous refinement. A database layer underpins the entire pipeline, storing both raw profile data and the outputs generated at each processing stage.

Users interact with the system through a web interface where they can submit an Instagram username for evaluation. The system then retrieves relevant profile metadata via the Instagram API, processes it through the pipeline, and returns a verdict—genuine or fake—along with a reporting button if the account is flagged as suspicious.

B. Algorithm Description

Step 1 – Data Collection: Profile data is gathered from the Instagram API, capturing attributes such as profile picture availability, username length, full name composition, number of posts, follower and following counts, account privacy status, and any linked external URLs.

Step 2 – Data Preprocessing: Raw data is cleaned to address missing values and standardised to bring all numerical features onto a comparable scale. This stage also removes redundant or irrelevant fields that are unlikely to contribute to classification.

Step 3 – Feature Extraction: Key discriminating features are derived from the preprocessed data. These include profile completeness scores, follower-to-following ratios, posting frequency, engagement rate relative to audience size, and the presence or absence of a profile photograph.

Step 4 – Fake Profile Detection: The extracted features are first evaluated against a set of rule-based thresholds (e.g., extremely low engagement combined with a missing profile image triggers an immediate flag). A Support Vector Machine (SVM) classifier then provides a probabilistic classification, with accounts whose scores exceed a defined threshold being labelled as fake.

Step 5 – Alert and Report: Users are notified of any suspicious classification and offered a one-click reporting mechanism that logs the account for review.

Step 6 – Continuous Improvement: Detection thresholds and model parameters are periodically reviewed and adjusted based on evolving data patterns and user feedback, ensuring the system remains effective against emerging fake account strategies.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

Step 7 – User Education: The interface incorporates short, contextual guidance notes that help users understand the warning signs of a fake profile, fostering greater awareness and more responsible social media use.

IV. RESULTS

Testing confirmed that the combined rule-based and machine learning pipeline is effective in distinguishing genuine accounts from fraudulent ones across a range of profile types. The system evaluates several profile attributes simultaneously—including the follower-to-following ratio, posting frequency, content repetition, and profile image authenticity—and produces a binary classification along with a confidence indicator.

When the username "justin95912" was submitted, the system retrieved the following data: username length of 11 characters; a full name composed of two words with a length of 12 characters; a description length of 100 characters; 65 posts; 8,194 followers; 2,680 accounts followed; the 'name equals username' flag set to FALSE; and the account marked as public. After processing these attributes, the classifier returned a 'Fake: TRUE' verdict, correctly identifying the account as inauthentic.

The interface presents the retrieved details in a clear tabular format and displays a 'Report User' button immediately below the verdict, enabling the user to submit a report with a single interaction. A confirmation message reading 'Report sent successfully' is displayed upon submission, closing the feedback loop.

Across a broader test set, the model demonstrated strong performance on standard classification metrics. Precision and recall values remained consistently high, indicating that the system generates relatively few false positives while capturing the majority of actual fake accounts. The F1-score reflected the balanced trade-off between these two objectives, confirming the model's suitability for real-world deployment.

V. COMPARISON ANALYSIS

Table 1 summarises seven detection approaches evaluated in terms of their underlying method, primary strengths, notable limitations, and the scenarios for which each is best suited.

Method	Approach	Strengths	Limitations	Best Used For
Rule-Based	Relies on predefined thresholds such as absent profile images or minimal user activity	Quick to deploy and produces fast outcomes	Cannot adapt to evolving deception tactics and may miss sophisticated fake profiles	Initial screening of candidate fake accounts
Machine Learning (SVM, RF)	Trains on labelled data to identify patterns linked to fake profiles	Adapts over time; effective at detecting subtle and non-obvious patterns	Requires a large, well-labelled dataset and significant training time	Large-scale detection in dynamic environments
Graph-Based	Analyses relationship structures, follower clusters, and community topology	Highly effective at uncovering coordinated bot networks	Demands extensive network data and considerable computing resources	Identifying bot farms and coordinated fake account clusters
Blockchain Reporting	Uses tamper-proof distributed records to log suspicious activity	Guarantees transparency and preserves the integrity of abuse reports	Technically complex; not suited to direct account classification	Long-term and secure documentation of reported incidents
Ensemble (Majority Voting)	Combines outputs from several classifiers before reaching a final verdict	High overall accuracy; reduces systematic errors of any single model	More resource-intensive and complex to maintain	Applications requiring high reliability and low error tolerance
Image Verification	Applies reverse image lookup and AI-based checks to identify stolen photographs	Effective at detecting image-based identity fraud	Limited effectiveness for private accounts or those using original images	Verifying profile photo authenticity

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

Method	Approach	Strengths	Limitations	Best Used For
API-Based Monitoring	Retrieves and analyses live profile data via social media APIs	Delivers real-time insights and up-to-date profile information	Constrained by API rate limits and platform privacy policies	Real-time profile surveillance and monitoring

Table 1: Comparative analysis of fake profile detection approaches.

VI. CONCLUSION

This research has demonstrated that automated detection of fake social media profiles is both technically feasible and practically valuable. By combining rule-based logic with a machine learning classifier trained on real Instagram profile data, the system achieves accurate identification of fraudulent accounts while minimising unnecessary disruption to legitimate users.

Beyond the technical outcomes, the project contributes to a wider goal of making online spaces more trustworthy. Fraudulent profiles carry genuine costs—financial losses, psychological harm, and the erosion of public confidence in digital communication—and tools that reduce their prevalence serve a meaningful social purpose. The integrated reporting mechanism strengthens this impact by converting detection into action: once a suspicious profile is identified, the pathway from discovery to removal is made as frictionless as possible.

Looking ahead, the system can be meaningfully extended in several directions. Real-time detection capabilities would allow threats to be neutralised before they cause harm, while cross-platform integration would extend protection to Facebook, Twitter, and other major networks. Automated escalation to platform moderators or relevant authorities could further reduce response times. The incorporation of identity verification tools—such as phone number validation or government ID checking—would add a further layer of authenticity assurance. Finally, embedding educational resources directly into the user interface could contribute to long-term digital literacy, helping users become more discerning participants in the social media ecosystem.

ACKNOWLEDGMENT

The authors wish to thank the faculty and research support staff at Vignan's Institute of Management and Technology for Women, Hyderabad, for their guidance and encouragement throughout this work. Gratitude is also extended to the open-source community whose tools and datasets made this research possible.

REFERENCES

[1] N. G. Kerrysa and I. Q. Utami, "Fake Account Detection in Social Media Using Machine Learning Methods: Literature Review," J. Inf. Syst. Eng. Bus. Intell., 2022.
[2] S. Deshmukh et al., "Fake Social Media Profile Detection and Reporting Using Blockchain Technology," Int. J. Adv. Res. Sci. Commun. Technol., 2023.
[3] A. Alharbi et al., "Social Media Identity Deception Detection: A Survey," ACM Comput. Surv., vol. 54, no. 3, pp. 1–36, 2021.
[4] D. R. Patil et al., "Detecting Fake Social Media Profiles Using the Majority Voting Approach," in Proc. Int. Conf. Comput. Intell. Data Sci., 2022, pp. 112–119.
[5] A. K. Singh and S. K. Singh, "Fake Profile Detection in Social Networks," Int. J. Comput. Appl., vol. 180, no. 12, pp. 18–23, 2018.
[6] S. K. Singh and S. Gupta, "Detecting Fake Social Media Profiles Using Machine Learning," in Proc. IEEE Int. Conf. Mach. Learn. Appl., 2020, pp. 87–93.
[7] N. R. Appini and V. B. Kumar, "Phishing URL Detection with Gradient Boosting Classifier," Commun. Appl. Nonlinear Anal., vol. 32, no. 3, 2025.
[8] M. Zabihimayvan and D. Doran, "Fuzzy Rough Set Feature Selection to Enhance Phishing Attack Detection," arXiv preprint arXiv:1910.10136, 2019.
[9] Ö. Kasim, "Automatic Detection of Phishing Pages with Event-Based Request Processing, Deep-Hybrid Feature Extraction and Light Gradient Boosted Machine Model," Telecommun. Syst., Springer, 2021.
[10] J. O. Ajayi and A. O. Adetunmbi, "Phishing Detection: Performance Evaluation of Both Ensemble and Classical Machine Learning Models," Int. J. Inf. Secur. Priv. Digit. Forensics, 2022.
[11] P. Prakash and M. Kumar, "PhishNet: Predictive Blacklisting to Detect Phishing Attacks," in Proc. Int. Conf. Ind. IoT Big Data Supply Chain, 2022, pp. 34–39.
[12] S. R. Curtis and P. Rajivan, "Phishing Attempts Among the Dark Triad: Patterns of Attack and Vulnerability," Comput. Hum. Behav., vol. 89, pp. 354–363, Oct. 2018.
[13] K. N. S. B. V. Manjushal and D. J. Kumari, "Detecting Phishing Links Analysis Using Machine Learning," Int. J. Innov. Multidiscip. Res., 2024.
[14] A. Alswailem et al., "Detecting Phishing Websites Using Machine Learning," in Proc. 2nd Int. Conf. Comput. Appl. Inf. Secur. (ICCAIS), 2019, pp. 1–6.
[15] J. Rashid et al., "Phishing Detection Using Machine Learning Technique," in Proc. 1st Int. Conf. Smart Syst. Emerg. Technol. (SMARTTECH), 2020, pp. 43–46.
[16] M. H. Alkawaz et al., "Detecting Phishing Websites Using Machine Learning," in Proc. 16th IEEE Int. Colloq. Signal Process. Appl. (CSPA), 2020, pp. 111–114.
[17] A. Razaque et al., "Detection of Phishing Websites Using Machine Learning," in Proc. IEEE Cloud Summit, 2020, pp. 103–107.
[18] T. Nagunwa, "Comparative Analysis of Nature-Inspired Metaheuristic Techniques for Optimizing Phishing Website Detection," MDPI Algorithms, 2024.
[19] D. Shanthi, "Smart Healthcare for Pregnant Women in Rural Areas," in Medical Imaging and Health Informatics, Wiley, ch. 17, pp. 317–334, 2022.
[20] D. Shanthi, R. K. Mohanty, and G. Narsimha, "Application of Machine Learning Reliability Data Sets," in Proc. 2nd Int. Conf. Intell. Comput. Control Syst. (ICICCS), 2018, pp. 1472–1474.