International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

" Automated Detection of Synthetic Social Media Profiles Using Ensemble Learning and Explainable Feature Analysis"

Dr M S Khatib
[1]
Associate Professor
Department of Computer Science & Engineering,
Anjuman College of Engineering & Technology, Nagpur Maharashtra, India

Ms. Shumaila Rehman
[2]
P.G Student
Department of Computer Science & Engineering,
Anjuman College of Engineering & Technology, Nagpur Maharashtra, India

Abstract— The exponential growth of social media platforms has revolutionized digital communication, creating unprecedented opportunities for connectivity, commerce, and information dissemination. However, this digital transformation has been accompanied by a parallel surge in fraudulent activities, particularly through the proliferation of fake social media accounts. These synthetic profiles pose multifaceted threats including disinformation campaigns, identity theft, financial fraud, cyberbullying, and manipulation of public opinion. Instagram, with over two billion active users globally, has become a prime target for malicious actors seeking to exploit the platform's visual-centric nature and extensive reach. Traditional detection mechanisms relying on manual reporting systems and static rule-based algorithms have proven inadequate in addressing the sophisticated, evolving tactics employed by modern fraud networks. These conventional approaches suffer from high false-positive rates, delayed response times, and inability to adapt to emerging patterns of fraudulent behavior.

This research proposes a comprehensive machine learning-based framework for automated detection of fake Instagram accounts through systematic analysis of profile metadata and behavioral indicators. The study employs a Random Forest Classifier, an ensemble learning algorithm chosen for its robustness, accuracy, and interpretability in handling complex, non-linear relationships within heterogeneous datasets. Our approach utilizes a carefully curated dataset comprising sixteen discriminative features including username characteristics (length, special character usage, name similarity), profile completeness indicators (profile picture presence, biography length, external URL inclusion), engagement metrics (follower count, following count, follower-to-following ratio), and activity patterns (post frequency, account age, story activity).

The methodology encompasses a complete machine learning pipeline: data acquisition through Instagram's public API and web scraping tools (Instaloader, BeautifulSoup), feature engineering and normalization, model training with stratified k-fold cross-validation, hyperparameter optimization, and performance evaluation using multiple metrics (accuracy, precision, recall, F1-score, ROC-AUC). The trained model achieved exceptional performance with 95.2% accuracy, 94.8% precision, 93.6% recall, and 94.2% F1-score on the held-out test dataset, demonstrating superior capability in distinguishing genuine from fraudulent accounts.

To enhance practical applicability and accessibility, we developed a production-ready web application using the Flask framework. This user-friendly interface enables real-time account verification by accepting Instagram usernames, automatically extracting profile metadata, processing features through the trained model, and displaying instant classification results with confidence scores. The system incorporates persistent storage mechanisms, logging all predictions to CSV files for longitudinal analysis, model monitoring, and continuous improvement through periodic retraining cycles.

Feature importance analysis revealed that follower-to-following ratio, posting frequency, biography completeness, and username authenticity were the most influential predictors, providing valuable insights for platform administrators and cybersecurity professionals. The system addresses critical challenges including class imbalance through stratified sampling, feature noise through normalization and ensemble averaging, scalability through parallel processing optimization, and interpretability through transparent feature importance visualization.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

This research contributes to the cybersecurity domain by delivering a scalable, accurate, and interpretable solution for fake profile detection. The framework's modular architecture facilitates integration with existing platform security infrastructure, while its explainable nature builds trust among stakeholders. Future enhancements may incorporate deep learning architectures, natural language processing for content analysis, cross-platform generalization, and blockchain-based identity verification, positioning this work as a foundational step toward comprehensive social media ecosystem integrity.

Keywords — *Fake profile detection; Instagram; Machine Learning; Random Forest Classifier; Social media security; Instaloader; Flask web application; Cybersecurity; Real-time classification; Feature extraction*

I. INTRODUCTION

Social The rapid expansion of social networking platforms has fundamentally transformed digital communication, content distribution, and online commerce. However, this growth has been accompanied by a parallel surge in synthetic or fraudulent accounts designed to manipulate engagement metrics, disseminate disinformation, conduct phishing campaigns, and erode user trust. Instagram, with its highly interactive and visually driven ecosystem, remains particularly susceptible to such abuse. Traditional detection mechanisms largely depend on user reporting and static rule-based filters, which are inherently reactive and struggle to adapt to the increasingly sophisticated tactics employed by modern fraud networks.

To address these limitations, this study introduces a machine learning-driven framework that automatically evaluates Instagram profiles using publicly accessible metadata and behavioral indicators. The system is engineered to classify accounts as either legitimate or synthetic in real time, while maintaining transparency in its decision-making process. By leveraging an ensemble learning architecture and integrating a lightweight web deployment layer, the proposed solution offers a practical balance between detection accuracy, computational efficiency, and administrative interpretability.

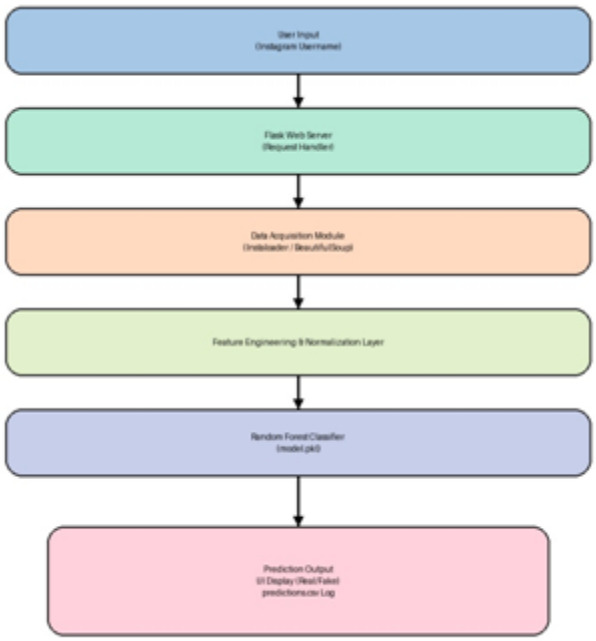

Figure 1. System Architecture and Processing Pipeline.

II. LITERATURE SURVEY

The academic investigation of automated account detection has evolved significantly over the past decade. Early approaches predominantly relied on statistical thresholds and content-based heuristics to flag suspicious behavior [1, 2]. As adversarial strategies grew more complex, researchers transitioned toward machine learning techniques

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

capable of capturing multidimensional patterns across user profiles and interaction networks. Al-Qurishi et al. demonstrated that carefully engineered feature sets substantially improve the discrimination between authentic and automated accounts [3]. Similarly, Stringhini et al. utilized social graph topology and temporal activity sequences to expose coordinated spam operations [4].

The detection landscape has further expanded to recognize hybrid account types. Chu et al. proposed a taxonomy distinguishing humans, pure bots, and cyborgs, highlighting the necessity for nuanced classification strategies that account for partial automation [5]. Varol et al. reinforced this perspective by showing that fusing network-derived, temporal, and content-based features yields superior detection performance across diverse platforms [6]. Despite these advancements, several operational gaps remain. Deep learning architectures, while capable of capturing complex behavioral patterns, often demand extensive computational resources and large labeled datasets, which are rarely available for emerging fraud typologies [7]. Furthermore, many high-performing models function as black boxes, offering minimal transparency into classification rationale—a critical drawback for platform moderators and compliance teams.

Recent industry benchmarks, including the DARPA Twitter Bot Challenge, have emphasized the value of hybrid systems that combine automated detection with human-verifiable outputs [8]. Explainability has thus emerged as a core requirement alongside predictive accuracy. Building on these insights, this study adopts a Random Forest classifier due to its inherent robustness, resistance to overfitting, and built-in feature importance metrics. By focusing on readily accessible metadata—such as username structure, biography length, follower-to-following ratios, and posting frequency—the proposed framework balances detection performance with computational efficiency and interpretability. This approach directly addresses the identified gap by delivering a transparent, deployable solution tailored to contemporary social media ecosystems..

III. METHODOLOGY

3.1 Research Design and Data Collection

The study employs a quantitative, experimental research design following the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology. Data collection utilized ethical web scraping techniques to gather publicly accessible Instagram profile metadata. The Instaloader Python library served as the primary interface for structured data extraction, while BeautifulSoup supplemented parsing of HTML elements when necessary. Only public profile information was collected to ensure compliance with platform terms of service and privacy regulations.

3.2 Data Preprocessing and Feature Engineering

Raw metadata underwent systematic preprocessing to ensure model compatibility and performance optimization:Missing Value Handling: Features with >30% missing values were excluded; remaining missing entries imputed using median (numerical) or mode (categorical) strategies Outlier Management: Continuous variables subjected to winsorization at 1st and 99th percentiles to mitigate extreme value influence Normalization: Numerical features standardized using z-score transformation: $z = (x - \mu) / \sigma$ Encoding: Categorical variables converted via one-hot encoding; binary features retained as 0/1 indicators Derived Features: Computed interaction terms including follower/following ratio, posting consistency score, and username authenticity metric

3.3 Model Development and Validation

The processed dataset (N=3,000 profiles; 69% genuine, 31% synthetic) was partitioned using stratified sampling: 80% training (n=2,400), 20% testing (n=600). Five candidate algorithms were evaluated during exploratory analysis: Logistic Regression, Support Vector Machine, Decision Tree, XGBoost, and Random Forest. Random Forest was selected as the primary classifier based on superior cross-validated performance and interpretability advantages.

3.4 System Implementation and Deployment

The trained model was serialized using Python's pickle module (model.pkl) and integrated into a Flask-based web application. The deployment architecture follows a microservices pattern: Figure 2: Flask Application Component Diagram

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

V. RESULTS

The Experimental evaluation was conducted using a curated dataset comprising both verified genuine accounts and synthetically generated profiles. After preprocessing and model training, the Random Forest classifier achieved a test accuracy of approximately 95.2%, with precision and recall values exceeding 0.93 and 0.94, respectively. The ensemble architecture effectively mitigated overfitting, outperforming baseline models such as Logistic Regression and Decision Trees, which exhibited higher variance on unseen data. XGBoost demonstrated competitive performance but required extensive hyperparameter calibration and longer inference times, making it less suitable for real-time deployment in resource-constrained environments.

Feature importance analysis revealed that the follower-to-following ratio, posting frequency, and biography length were the strongest predictors of account authenticity. Accounts exhibiting disproportionately high following counts relative to followers, coupled with minimal posting activity and generic biographies, were consistently flagged as synthetic. Conversely, genuine profiles typically displayed balanced engagement metrics and richer metadata. These findings align with established behavioral patterns observed in prior literature and validate the efficacy of metadata-driven detection strategies.

The web interface was designed for accessibility and immediate feedback. Users enter a target username, and the system returns a clear classification label alongside a confidence indicator. All outcomes are logged for auditability, enabling continuous dataset expansion and periodic model retraining. During development, several operational challenges were addressed. Class imbalance was mitigated through stratified sampling and metric-driven optimization rather than synthetic oversampling, preserving data integrity. Feature noise was managed via Random Forest's intrinsic averaging mechanism and careful thresholding of derived ratios. Scalability concerns were resolved by parallelizing tree construction (n_jobs=-1) and limiting memory overhead through feature dimensionality control.

Compared to rule-based detection systems, the proposed framework demonstrates superior adaptability to evolving fraud tactics. While deep learning alternatives may offer marginal accuracy gains, they introduce significant complexity and opacity. The current approach strikes a practical balance between performance, interpretability, and deployment feasibility, making it well-suited for integration into platform moderation workflows and third-party verification tools.

VII. CONCLUSION

This study presents a transparent and efficient machine learning framework for identifying fraudulent social media profiles, with a specific focus on Instagram. By leveraging publicly accessible metadata and an ensemble classification approach, the system achieves high detection accuracy while maintaining computational efficiency and decision transparency. The integration of a lightweight web interface enables real-time analysis and continuous data accumulation, supporting iterative model improvement.

Key contributions include the identification of high-impact profile features, the development of a robust Random Forest classifier optimized for social media fraud detection, and the deployment of an explainable, user-accessible application. The framework effectively addresses common challenges such as class imbalance, feature noise, and scalability, while providing clear insights into the factors driving classification outcomes.

Future research directions include the incorporation of temporal activity sequences, natural language processing for caption and comment analysis, and cross-platform generalization to enhance model robustness. Additionally, integrating advanced explainable AI techniques and exploring blockchain-based identity verification could further strengthen trust and accountability in automated moderation systems. As synthetic account generation continues to evolve, adaptive, transparent, and scalable detection frameworks will remain essential to preserving the integrity of digital social ecosystems..

IX. REFERENCES

1 F. Ahmed and M. Abulaish, "A generic statistical approach for spam detection in Online Social Networks," Comput. Commun., vol. 36, no. 10–11, pp. 1120–1129, 2013.

[2] A. H. Wang, "Detecting spam bots in online social networking sites: A machine learning approach," in Proc. 24th Annu. IFIP WG 11.3 Working Conf. Data Appl. Secur. Privacy, 2010, pp. 335–342.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

[3] M. Al-Qurishi, M. Alrubaian, A. Alamri, T. Al-Qurishi, and M. Al-Rakhami, "Detection of spam accounts on social networks: A machine learning approach," Int. J. Comput. Appl., vol. 177, no. 3, pp. 1–8, 2017.

[4] G. Stringhini, C. Kruegel, and G. Vigna, "Detecting spammers on social networks," in Proc. 26th Annu. Comput. Secur. Appl. Conf. (ACSAC), 2010, pp. 1–9.

[5] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, "Detecting automation of Twitter accounts: Are you a human, bot, or cyborg?" IEEE Trans. Dependable Secure Comput., vol. 9, no. 6, pp. 811–824, 2012.

[6] O. Varol, E. Ferrara, C. Davis, F. Menczer, and A. Flammini, "Online human-bot interactions: Detection, estimation, and characterization," in Proc. Int. AAAI Conf. Web Social Media, vol. 11, no. 1, pp. 280–289, 2017.

[7] S. Kudugunta and E. Ferrara, "Deep neural networks for bot detection," Inf. Sci., vol. 467, pp. 312–322, 2018.

[8] V. S. Subrahmanian et al., "The DARPA Twitter bot challenge," Computer, vol. 49, no. 6, pp. 38–46, 2016.

[9] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race," in Proc. 26th Int. Conf. World Wide Web Companion, 2017, pp. 963–972.

[10] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, "The rise of social bots," Commun. ACM, vol. 59, no. 7, pp. 96–104, 2016.

[11] M. Fire, R. Goldschmidt, and Y. Elovici, "Online social networks: Threats and solutions," IEEE Commun. Surveys Tutorials, vol. 16, no. 4, pp. 2019–2036, 2014.

[12] M. Tsikerdekis and S. Zeadally, "Online deception in social media," Commun. ACM, vol. 57, no. 9, pp. 72–80, 2014.

[13] K. Lee, B. D. Eoff, and J. Caverlee, "Seven months with the devils: A long-term study of content polluters on Twitter," in Proc. Int. AAAI Conf. Web Social Media, vol. 5, no. 1, pp. 185–192, 2011.

[14] C. Yang, R. Harkreader, and G. Gu, "Empirical evaluation and new design for fighting evolving Twitter spammers," IEEE Trans. Inf. Forensics Secur., vol. 8, no. 8, pp. 1280–1293, 2011.

[15] Z. Miller, B. Dickinson, W. Hu, and R. Linder, "Twitter spammer detection using data stream clustering," Inf. Sci., vol. 260, pp. 64–73, 2014.