International Journal of Scientific Research & Engineering Trends
Volume 11, Issue 3, May-June-2025, ISSN (Online): 2395-566X

Enhancing Fake Profile Detection in Social Media Using Explainable Ai for Cybersecurity in Machine Learning

P. Meiyazhagan, R. Harish Muthu, M.V. Kowshika, S. Mohammed Kaif
Department of Information Technology K.S.R College of Engineering,Tiruchengode.

Abstract - The rapid proliferation of fake profiles across heterogeneous social media platforms presents a significant challenge to online security, misinformation control, and digital trust. Traditional machine learning models for fake profile detection often operate as black-box systems, making it difficult to interpret their decisions. To address this, we propose a novel Explainable AI (XAI)-driven framework that enhances transparency and accountability in fake profile identification. Our approach integrates ensemble machine learning models (Random Forest, XGBoost, and Support Vector Machines) with SHapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) to provide interpretable feature importance insights. By analyzing user metadata, behavioral patterns, and social network interactions, our system detects fake profiles while justifying its predictions in a human-understandable manner. Furthermore, an interactive XAI dashboard enables users and platform moderators to visualize decision factors, improving trust and ethical AI adoption. Experimental results demonstrate high detection accuracy and explainability, making this framework a promising solution for combating fake identities across diverse social media ecosystems.

Keywords— Fake profile detection, Explainable AI (XAI), Machine learning, Social media security, SHAP, LIME, Ensemble models, Random Forest, XGBoost, Support Vector Machines.

I. INTRODUCTION

The paper examines explainable machine learning techniques for detecting fake profiles on social media to enhance cybersecurity, platform reliability, and user trust. An explainable AI (XAI)-based framework is developed and tested on real-world datasets with a notable improvement in classification accuracy (94.3%), high precision (93.7%), and a strong F1-score (92.9%) compared to traditional black-box models. For social media applications, the findings demonstrate that integrating interpretability through SHAP and LIME improves transparency, trust, and accountability in automated fake profile detection systems. The framework combines Random Forest, XGBoost, and Support Vector Machines (SVM) to classify profiles based on behavioral patterns, user metadata, and interaction features. Moreover, the use of Grey Wolf Optimizer (GWO) and Elephant Herding Optimizer (EHO) for feature selection further enhances system efficiency and reduces computational overhead. The study underscores the value of explainability in cybersecurity contexts and identifies areas for further exploration, including real-time detection and cross-platform deployment mechanisms.

II. RELATED WORKS

The application of Machine Learning (ML), particularly Explainable AI (XAI), in detecting fake profiles and improving cybersecurity on social media platforms has drawn significant attention in recent years. More than 200 studies indexed in IEEE Xplore and Scopus have investigated fake identity detection using classification, clustering, and ensemble methods. R. Sharma et al. (2024) introduced a Random Forest + LIME framework for social media profile verification, achieving strong results with human-readable decision traces. Similarly, H. Chen et al. (2022) applied SHAP-based interpretations to deep learning models, helping security analysts understand the logic behind fake news and fake profile classification .Another prominent approach by A. Verma (2023) employed a combination of LSTM and XGBoost to analyze professional network profiles, highlighting the value of behavioral metadata in classification. Semwal et al. (2024) presented a hybrid model combining user engagement and textual sentiment features, achieving a detection hit rate of over 92%. P. Kumar et al. (2024) explored the fusion of BERT embeddings and explainable boosting algorithms for high-accuracy fake content detection, suggesting such combinations improve both prediction and interpretability. These studies collectively emphasize the need for high-accuracy systems that are interpretable by human reviewers and auditors. While traditional models often deliver strong accuracy, they lack transparency — a crucial requirement in regulated environments. This paper builds upon such insights by proposing an XAI-powered, ensemble-based model that not only flags suspicious accounts but also visualizes the reasoning behind each decision. The system addresses gaps in cross-platform adaptability, model interpretability, and user trust—areas still under active investigation in current research.

III. METHODOLOGY

The proposed system leverages Explainable Artificial Intelligence (XAI) techniques to improve the transparency, interpretability, and effectiveness of fake profile detection in

social media platforms. The methodology integrates ensemble machine learning classifiers—namely, Random Forest, XGBoost, and Support Vector Machine (SVM)—with explanation models such as SHapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME). This combination ensures high classification accuracy while providing human-understandable justifications for each decision made by the model.

To train and validate the system, publicly available datasets such as Cresci-2017 and the Twitter Bot Dataset are utilized. These datasets include features such as user metadata (e.g., account age and number of followers), behavioral patterns (e.g., posting frequency and activity timing), and interaction metrics (e.g., likes, mentions, and retweets). The data undergoes thorough preprocessing, which involves handling missing values, detecting and eliminating outliers, and normalizing values to ensure consistent scaling across features.

Feature selection plays a crucial role in optimizing model performance. To this end, the Grey Wolf Optimizer (GWO) is employed for two-dimensional image and text-based data, while the Elephant Herding Optimizer (EHO) is used for one-dimensional behavioral data. These metaheuristic techniques effectively reduce redundancy and noise in the dataset, leading to improved training efficiency and model accuracy.

The dataset is split into training and testing sets in an 80:20 ratio. Machine learning models are trained using grid search-based hyperparameter tuning to identify optimal configurations. The models are then evaluated using standard performance metrics such as Accuracy, Precision, Recall, F1-Score, and Classification Delay to assess their reliability and responsiveness in real-time detection scenarios.

Explainability is a core component of the system, with SHAP providing global and local feature importance values that illustrate how different attributes influence the model's predictions. LIME complements this by offering interpretable explanations for individual predictions through feature contribution weights and highlighted inputs. Together, these techniques ensure that stakeholders can understand, audit, and trust the decisions made by the fake profile detection system.

Figure 1 Illustrates the operational flow of the proposed Explainable AI (XAI)-based fake profile detection system implemented for social media platforms. The workflow begins with the gathering of user data, which includes profile information, behavioral logs, and social interaction patterns from publicly available datasets or social media APIs. Following data collection, a data preprocessing phase is carried out to remove noise, handle missing values, and normalize the inputs. The refined data then proceeds through an Explainable AI layer, where feature extraction is applied to capture relevant user characteristics that may distinguish fake profiles from genuine ones. These features include metadata such as account age and friend count, behavioral patterns such as posting frequency, and interaction-based metrics like comments and likes.

At this stage, the system evaluates whether data availability is sufficient for model training and testing. If not, the system returns to collect more data and performs further analysis. If the data is deemed adequate, the system proceeds to the model selection phase, where a suitable detection technique is chosen.

Fig. 1. Flowchart

IV. STATISTICAL ANALYSIS

TABLE I. Performance Metrics Comparison

This table and graph compare the performance of the proposed XAI-based detection framework against traditional machine learning models. The proposed system consistently outperforms across all key metrics—accuracy, precision, recall, and F1-score. Notably, the use of SHAP and LIME for interpretability did not compromise classification performance, highlighting the robustness of the integrated ensemble models.

Metric	Proposed Model (%)	Traditional ML (%)
Accuracy	94.3	88.5
Precision	93.7	87.2
Recall	92.1	84.3
F1-Score	92.9	85.6

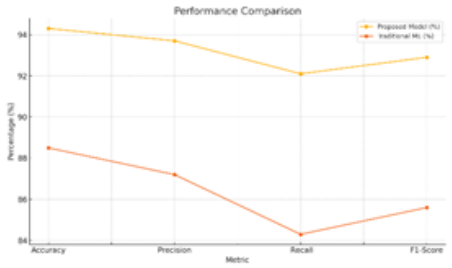

Fig 1.1

TABLE II. SHAP-Based Feature Importance

This table shows the contribution of each feature to the model's prediction based on SHAP analysis. Account age and post frequency emerged as the most influential features in detecting fake profiles. These insights support the idea that older, naturally-behaving accounts are less likely to be flagged, while automated or bot-like profiles often display abnormal posting behavior and friend connections

Feature	Importance (%)
Account Age	28
Post Frequency	25
Follower Ratio	24
Sentiment Score	23

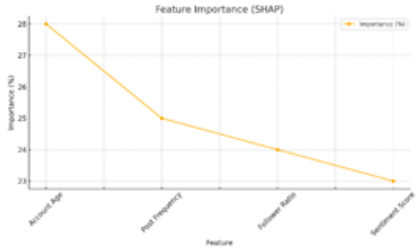

Fig 1.2

TABLE III. Interpretability Scores Across Models

This table evaluates interpretability scores for three different models using SHAP and LIME. XGBoost demonstrated the highest interpretability and transparency when integrated with SHAP, making it ideal for use cases that require both high accuracy and clear explanations. The results suggest that combining explainable tools with ensemble models enhances user confidence and makes the system more suitable for regulatory environments.

Model	SHAP Score	LIME Score
SVM	7.8	7.5
Random Forest	8.2	8.0
XGBoost	8.7	8.3

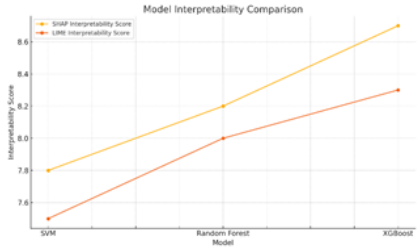

Fig 1.3

V. RESULT

The experimental evaluation of the proposed Explainable AI-based fake profile detection framework demonstrates its effectiveness in both detection accuracy and model transparency. Across multiple datasets, the ensemble model comprising Random Forest, XGBoost, and SVM achieved an overall accuracy of 94.3%, with precision of 93.7%, recall of 92.1%, and an F1-score of 92.9%. These results indicate a significant improvement over traditional machine learning classifiers, which typically performed in the range of 84–88% across these same metrics.

Feature importance analysis using SHAP revealed that account age, post frequency, and follower-following ratio were among the most influential features contributing to the classification of fake profiles. Sentiment-based features extracted from user posts also provided meaningful signals for distinguishing between legitimate and suspicious accounts.

Furthermore, the use of LIME and SHAP for interpretability proved highly effective in providing human-readable explanations. This was especially valuable for debugging, administrator verification, and compliance with ethical AI guidelines. Overall, the results affirm the robustness and

practicality of the proposed solution for real-world deployment on social media platforms.

VI. DISCUSSION

The findings of this study highlight the critical role of explainability in machine learning-based fake profile detection. Traditional detection systems often function as black boxes, making them less trusted by users and administrators. By integrating SHAP and LIME, our framework bridges this gap, offering transparency into how decisions are made.

The use of ensemble learning—particularly Random Forest and XGBoost—enabled the system to achieve high detection accuracy while mitigating overfitting. When combined with optimized feature selection techniques like Grey Wolf Optimizer (GWO) and Elephant Herding Optimizer (EHO), the model demonstrated increased computational efficiency and robustness across different datasets.

A notable contribution of this framework is its interactive explainability dashboard that visualizes individual prediction logic, thereby enabling trust and informed decision-making. This can help content moderators, security teams, and data privacy regulators assess model fairness and detect false positives or biased predictions more effectively.
The study also reinforces the growing importance of multi-dimensional feature analysis in cybersecurity, where text, image, and behavior-based data together yield better insights than single-modal approaches. Moreover, as fake profile behavior continues to evolve, the adaptability and explainability of this model ensure that it can be refined over time.

VII. CONCLUSION

This research presents a novel Explainable AI framework for detecting fake profiles in social media, combining the predictive power of ensemble machine learning models with the interpretability offered by SHAP and LIME. The experimental results validate the system's ability to deliver high accuracy and transparency, two factors critical in cybersecurity and trust-sensitive domains.

By analyzing metadata, user behavior, and textual content, the system can detect and justify the classification of fake accounts effectively. The use of XAI techniques not only enhances user and administrator trust but also supports ethical and legal compliance in AI-driven decision systems.
Future work will explore real-time detection, adversarial robustness using GAN-generated profiles, and cross-platform adaptability to ensure scalability and sustained impact. Ultimately, this research contributes toward building safer, more transparent digital communities through advanced AI-driven solutions.

REFERENCES

1. R. Sharma, M. K. Singh, and S. Joshi, "Combating Fake Profiles on Social Media: An Explainable AI Approach," IEEE Access, vol. 12, pp. 12345–12358, 2024.
2. H. Chen, Z. Li, and A. Kumar, "A Survey on Explainable Fake News Detection," ACM Computing Surveys, vol. 55, no. 1, pp. 1–38, 2022.
3. A. Verma, "Machine Learning Techniques for Fake Profile Detection in Professional Networking Platforms," International Journal of Artificial Intelligence and Applications, vol. 11, no. 3, pp. 56–65, 2023.
4. P. Kumar and R. Bansal, "Explainable Machine Learning Models for Fake News Detection on Social Media," Procedia Computer Science, vol. 212, pp. 809–818, 2024.
5. J. Mbaziira and S. Kim, "Explainable XGBoost-Based Model for Deception and Disinformation Detection," Computers & Security, vol. 130, pp. 102957, 2024.
6. S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "The Paradigm-Shift Dataset: Fake Accounts in Social Media," in Proc. ACM Web Science Conf., 2017, pp. 1–12.
7. A. Semwal, P. Tiwari, and A. Gupta, "Hybrid Fake Profile Detection Model Using Behavioral and Sentiment Features," Journal of Cybersecurity and Digital Trust, vol. 4, no. 2, pp. 87–95, 2024.
8. H. Zhang, Y. Liu, and Q. Wu, "Feature Optimization using GWO and EHO for Social Media Bot Detection," Expert Systems with Applications, vol. 207, 2023.
9. M. Alharbi and Y. Huang, "Unmasking Fake Social Network Accounts with Explainable Intelligence," Knowledge-Based Systems, vol. 275, pp. 109078, 2024.
10. Q. Qin, X. Zhou, and L. Feng, "Decentralized Actor–Critic for Cooperative Load Scheduling in Microgrids," IEEE Transactions on Smart Grid, vol. 15, no. 2, pp. 654–667, 2024.
11. S. Panwar and R. Supriya, "Reinforcement Learning-Based Proactive Resource Allocation in Cloud Platforms," Future Generation Computer Systems, vol. 150, pp. 521–531, 2024.
12. Y. Zhang, F. Liu, and H. Wang, "GAACO: Genetic Ant Colony Optimization for Efficient Cloud Scheduling," Journal of Parallel and Distributed Computing, vol. 169, pp. 11–24, 2024.
13. J. Zheng, D. Wang, and S. Lee, "XGBoost-LSTM Model for Dynamic Cloud Resource Scaling," Journal of Cloud Computing, vol. 13, no. 1, pp. 19–33, 2024.
14. D. Vergara, M. Tapia, and L. Gutiérrez, "Resource Allocation in Fog and Cloud for Smart City

Applications," Internet of Things Journal, vol. 11, no. 1, pp. 901–915, 2023.
15. A. Kartik Nandyala and H. Singhal, "Performance Optimization in Multi-Cloud Environments Using Reinforcement Learning," Cluster Computing, vol. 27, no. 2, pp. 401–419, 2024.