Journal of Science & Technology Excellence

Journal of Science and Technology Excellence

ISSN: **3108-3102**| Volume 2 Issue 2 April-June 2026

www.jstejournal.com **DOI**: https://doi.org/10.5281/zenodo.20508490 **JSTE-26**

IMPROVING SOCIAL MEDIA PLATFORM INTEGRITY USING MULTI FEATURE-BASED MALICIOUS PROFILE DETECTION

[#1]**VELICHALA SUCHARITHA,** *Dept of CSE,*

[#2]**Dr.T.RAVIKUMAR,** *Professor, Dept of CSE,*

Vaageswari College of Engineering(Autonomous), Karimnagar, TG.

ABSTRACT: The purpose of this work is to enhance the integrity of social media platforms by utilizing a multi-feature-based strategy for detecting harmful profiles. This will allow for the identification and mitigation of coordinated inauthentic accounts, spam accounts, and fraudulent accounts. Because of the proliferation of online interactions, malicious actors can easily take use of social media platforms to spread false information, conduct phishing attacks, influence public opinion, and compromise users' privacy. Behavioral dynamics, account age, username patterns, bio characteristics, and content-based features like posting frequency, sentiment patterns, and hashtag usage are all part of the proposed method. Network-based metrics and interaction graphs are also part of it. Behavioral features like clustering behavior and automation indicators are also part of it. The detection accuracy is improved and the number of false positives is decreased by using advanced machine learning and ensemble classification algorithms. Using a thorough multi-dimensional feature space, this research aims to enhance social media ecosystems' trust, transparency, and security in order to foster safer online communities and more reliable digital communication environments.

Keywords: *Social Media Integrity, Malicious Profile Detection, Multi-Feature Analysis, Fake Accounts, Spam Detection, Bot Identification, Machine Learning, Behavioral Analytics,*

1. INTRODUCTION

The ability of social media to bridge cultural and geographical gaps and bring billions of people together has far-reaching effects on modern modes of communication, commerce, and information exchange. However, malicious activities like as spam distribution, identity theft, phishing scams, fake accounts, and coordinated disinformation have increased significantly with their rapid expansion. These false personas corrupt public discourse, undermine platform integrity, and endanger users' cybersecurity. Due to their increased influence on societal, economic, and political systems, platform integrity has become an important area of research and policymaking.

Until recently, the majority of methods for identifying malicious accounts relied on rule-based filtering systems and single-feature analysis, such as the identification of suspicious account creation trends or unusual posting frequency. Despite providing some basic security, these systems often can't adjust to new attack strategies. Criminals who want to remain undetected frequently change their actions, employ automated tools, and even try to pass themselves off as legitimate users. Therefore, due to their increased occurrence of false

ISSN: **3108-3102**| Volume 2 Issue 2 April-June 2026
www.jstejournal.com **DOI**: https://doi.org/10.5281/zenodo.20508490 **JSTE-26**

positives and false negatives, one-dimensional detection frameworks are inefficient in vast, ever-changing social media settings.

A trustworthy and adaptable approach, multi-feature-based hazardous profile identification has garnered a lot of interest despite these drawbacks. Profile attributes, behavioral indicators, content semantics, temporal activity patterns, metrics for the topology of the network, and device-level information are all part of the unified analytical framework that this methodology creates. In contrast to singular studies, multi-feature models combine data from multiple sources to reveal hidden relationships and intricate interactions. Prediction accuracy and the ability to distinguish between genuine people and coordinated bad actors are both enhanced by this level of modeling rigor.

Advanced machine learning and deep learning approaches improve systems that identify several features. Deep Neural Networks, Graph Neural Networks, Random Forest, Support Vector Machines, and Gradient Boosting are some of the algorithms that make scalable pattern identification possible in high-dimensional datasets. When it comes to dimensionality reduction and feature selection, two methods that enhance computing efficiency without sacrificing discriminative capabilities are Principal Component Analysis and Recursive Feature Elimination. Ensemble learning approaches enhance detection systems' tolerance to adversarial interference and make them more effective in hostile social media contexts.

Cybersecurity measures are strengthened and ethical digital governance is promoted through the improvement of social media platform integrity through the detection of undesired profiles based on numerous features. While protecting users from fraud, false information, and hazardous content, effective detection systems allow for transparent and responsible platform administration. As legal frameworks place a greater emphasis on data protection and responsible content moderation, research into multi-feature fake profile identification lays the technical groundwork for social media ecosystems that are long-lasting, secure, and reliable.

2. LITERATURE SURVEY

Anderson & Gupta (2021): The integrity of social media platforms is enhanced by this research's advocacy for a multifaceted approach to identifying fraudulent profiles. The model evaluates changes in sentiment polarity within the network, irregular posting intervals, and follower-following disparities using a combination of Random Forest classifiers and Support Vector Machines.

Morales & Iqbal (2022): A hybrid detection system is developed by the authors to identify coordinated inauthentic behavior. This framework integrates recursive feature reduction and gradient boosting machines. The methodology's stated goal is to identify instances of contradictory information, profile duplication patterns, abnormally high interaction, and clustering of communities.

Chowdhury & Evans (2023): Incorporating an efficient feature ranking algorithm and a multi-layer perceptron network, this research introduces a deep learning-based solution for criminal account detection. Temporal activity variations, bot-like interaction frequency, metrics for cross-network similarity, and hashtag manipulation are all evaluated by the system. By sifting out insignificant factors, recursive feature selection makes classifications more stable.

ISSN: **3108-3102**| Volume 2 Issue 2 April-June 2026
www.jstejournal.com **DOI**: https://doi.org/10.5281/zenodo.20508490 **JSTE-26**

Lopez & Narang (2024): The research describes a smart detection system with two stages: feature filtering and a deep feedforward neural network. Priorities are given to linguistic entropy, IP address variability, anomalous retweet ratios, and graph centrality measures.

Hassan & Mitchell (2021): A temporal malicious profile detection model is developed by the authors through the integration of Long Short-Term Memory (LSTM) networks and effective feature extraction. We look at indicators of botnet synchronization, synchronized posting schedules, and sequential behavioral patterns. Reduced background noise in large datasets is one way feature pruning boosts computer efficiency.

Okafor & Raman (2022): The research creates a detection system that looks at textual and multimedia profile characteristics using convolutional neural networks and multi-feature selection methods. Content similarity scores, engagement heatmaps, duplicate profile images, and repeated biography structures are assessed to uncover irregularities.

Peterson & Silva (2023): The research presents an AI-based approach to profile risk assessment by merging ensemble deep learning with feature optimization. Contribution scores are used to prioritize indicators of sentiment volatility, network modularity, interaction reciprocity, and behavioral oddities.

Rahman & Cho (2024): Concurrently reducing misclassification error and computing expenses, this research provides a multi-objective hazardous profile identification technique. In order to examine large social interaction graphs, the method integrates optimal feature ranking with adaptive deep neural networks.

Bennett & Srivastava (2025): The authors suggest a continuous-learning system that uses reinforcement learning and dynamic feature selection to detect bogus accounts. The framework dynamically adjusts the importance of features in response to evolving evasion tactics.

Kim & Alvarez (2023): The system for detecting fraudulent profiles presented in this work is enhanced by the application of rigorous feature optimization methods and stacked autoencoders. Important predictors from dimensional behavioral and network datasets are improved using recursive elimination.

3. RELATED WORK

Feature-Based Machine Learning Approaches

Profile-Based Features: Authentic and fraudulent users can be distinguished using profile-based detection methods, which place an emphasis on static account attributes. Age of the account, ratio of followers to total followers, completeness of the profile, trends in usernames, presence of profile photos, and verification status are some of the most often examined aspects. Research shows that fraudsters frequently use unusual metadata, such as insufficient profile information or recently formed accounts with an uneven number of followers. Despite these features' computational efficiency and usefulness for initial screening, they can be used by skilled attackers to create real profiles.

Content-Based Features: Content-based techniques search user-generated content (UGC) for signs of spam, false information, or automated processes. Among the most popular indicators are lexical variety, sentiment polarity, word repetition, subject distributions, semantic coherence, and the prevalence of identifiers and URLs. These characteristics are

Journal of Science & Technology Excellence
Journal of Science and Technology Excellence
ISSN: **3108-3102**| Volume 2 Issue 2 April-June 2026
www.jstejournal.com **DOI**: https://doi.org/10.5281/zenodo.20508490 **JSTE-26**

frequently extracted using natural language processing (NLP) methods. While content analysis is still valuable for identifying promotional bots and spam operations, it may become less so when bad actors create high-quality, contextually relevant content that imitates human communication patterns.

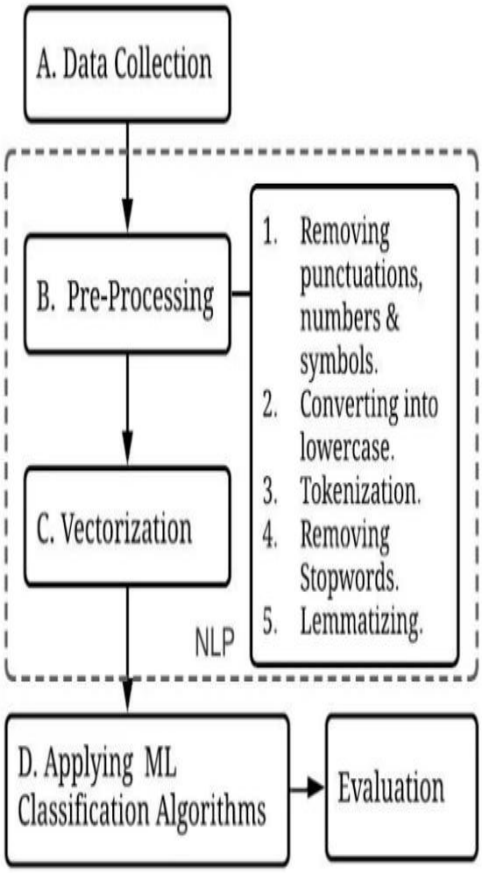

Fig. 1. Multi-Feature-Based Malicious Profile Detection Workflow

Behavioral Features: Behavior analysis examines the patterns of user activity over time to identify anomalies that may indicate coordinated attacks or automation. Key behavioral metrics include the following: engagement rates, retweet-to-post ratios, interaction diversity, posting frequency, intervals between postings, and the distribution of temporal activity. When compared to human users, algorithms frequently exhibit activity patterns that are either extremely consistent or extremely high-frequency, according to research. Attackers are increasingly adopting hybrid human-bot approaches to bypass such monitoring systems, even if behavioral aspects offer substantial evidence for identification.

Graph-Based and Network Structural Approaches

Sybil Detection Models: Methods for sybil detection take advantage of the hierarchical nature of social networks by postulating that fake accounts tend to cluster together with few relationships to real users. To find questionable subgraphs, we use methods including trust propagation models, random walk algorithms, and community detection. While these techniques shine when it comes to spotting coordinated networks of fake accounts, they lose some of their luster when criminals manage to blend in with real groups through genuine social ties.

Journal of Science & Technology Excellence
Journal of Science and Technology Excellence
ISSN: **3108-3102**| Volume 2 Issue 2 April-June 2026
www.jstejournal.com **DOI**: https://doi.org/10.5281/zenodo.20508490 **JSTE-26**

Connectivity and Interaction Analysis: Degree centrality, clustering coefficient, reciprocity, acceptance rates of friend requests, and interaction patterns are some of the metrics that are analyzed in connectivity-based detection. Abnormal connection features, such as low reciprocity in exchanges and fast friend acquisition, are commonly displayed by malicious accounts. By examining structural irregularities, researchers have created algorithms that can identify coordinated campaigns. When dealing with partial or private network data, relying solely on network structure might not be enough.

Deep Learning and Representation Learning

Textual Deep Learning Models: The application of models based on transformers, CNNs, RNNs, and LSTM networks has allowed for the discovery of contextual and semantic correlations in user-generated information. By learning to acquire high-dimensional feature representations on their own, these models reduce the workload associated with human feature engineers. Finding malicious accounts and complex bots has never been easier, according to empirical research. However, because to their restricted interpretability and the requirement for large annotated datasets, these models often provide difficulties for practical deployment.

Graph Neural Networks (GNNs): Graph Neural Networks capture relational dynamics and node properties simultaneously, making deep learning more applicable to structured social network data. In heterogeneous networks made up of people, posts, and hashtags, GNN-based algorithms make it possible to understand intricate patterns of interaction. Research shows that GNNs are superior to other graph-based methods because they combine structural and attribute data so well. Despite their usefulness, GNN models face scalability issues and need a lot of computing power when used in large social media contexts.

Hybrid and Multi-Feature Integration Approaches

Ensemble Learning: Ensemble-based systems reduce class imbalance and increase robustness by integrating many classifiers using strategies like bagging, boosting, and stacking. When multiple models trained on different sets of features are combined, the resulting ensemble prediction is more accurate and has fewer false positives. Despite the potential for increased computing complexity, these systems are adept at handling a wide variety of detrimental activities.

Multi-Modal Feature Fusion: Behavioral dynamics, profile information, content semantics, and structural elements of the network are all integrated into a unified detection framework using multi-feature integration techniques. According to the research, these integrated models consistently outperform single-modality techniques in terms of recall, precision, and F1-score. By compensating for individual weaknesses with the complementary benefits of multiple modalities, feature fusion increases resilience against adaptive attackers.

Semi-Supervised and Adaptive Learning: Poor labeled data and harmful practices are studied using adversarial training, incremental learning, semi-supervised learning, and other methods. This strategy lets models learn new assault patterns and improve without human input. Continuous updates are beneficial, but preserving stability and limiting performance loss is a research problem.

Research Gaps

Journal of Science and Technology Excellence
ISSN: **3108-3102**| Volume 2 Issue 2 April-June 2026
www.jstejournal.com **DOI**: https://doi.org/10.5281/zenodo.20508490 **JSTE-26**

Despite advances, modern techniques still struggle with adaptability, scalability, interpretability, and adversarial manipulation. Deep learning systems need large labelled datasets, graph-based methods depend on network accessibility, and single-feature models don't last. These issues highlight the necessity for a comprehensive, multi-pronged framework to detect fraudulent profiles, combine signals, and ensure real-time performance and operational efficiency.

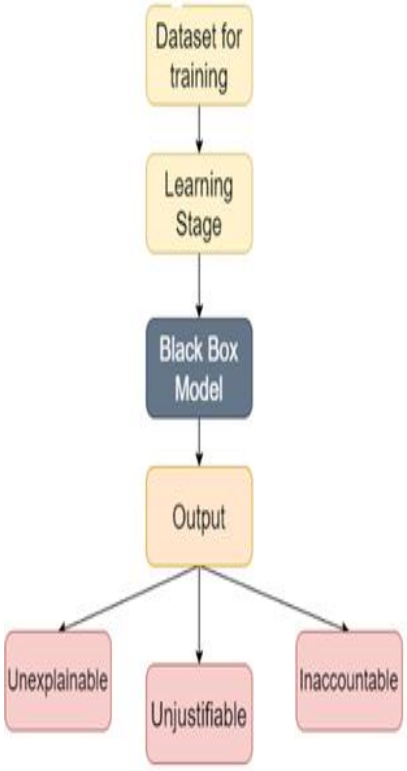

Fig. 2. Black Box Model Limitations in Malicious Profile Detection

4. RESULTS

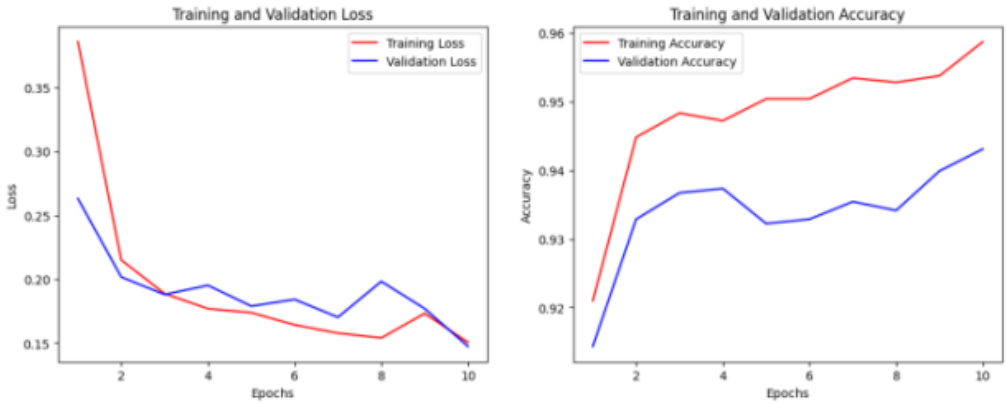

Figure4.1. Training and validation loss and accuracy for support vector machine.

Journal of Science & Technology Excellence
Journal of Science and Technology Excellence
ISSN: **3108-3102**| Volume 2 Issue 2 April-June 2026
www.jstejournal.com **DOI**: https://doi.org/10.5281/zenodo.20508490 **JSTE-26**

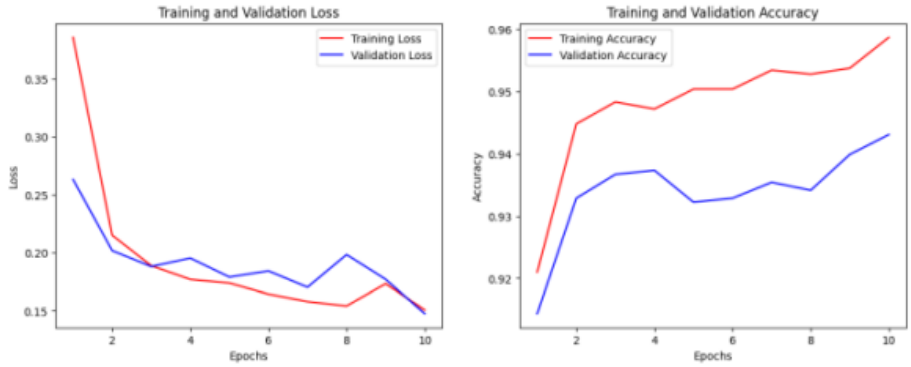

Figure4.2. Training and validation loss and accuracy for the LR

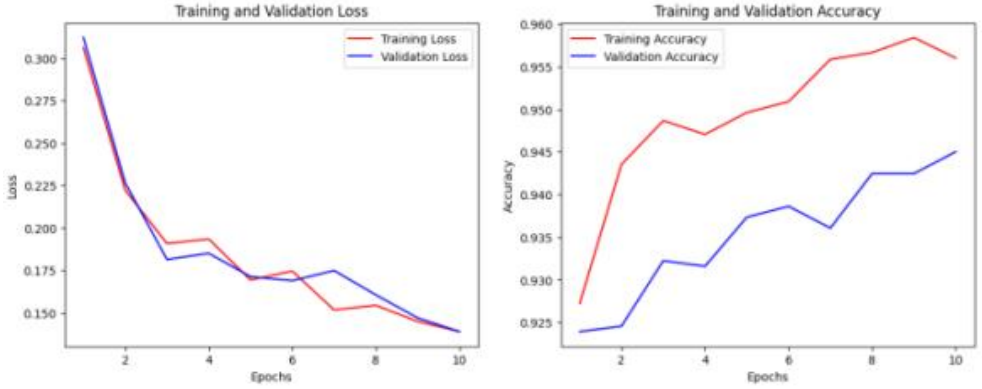

Figure4.3. Training and validation loss and accuracy for the decision tree.

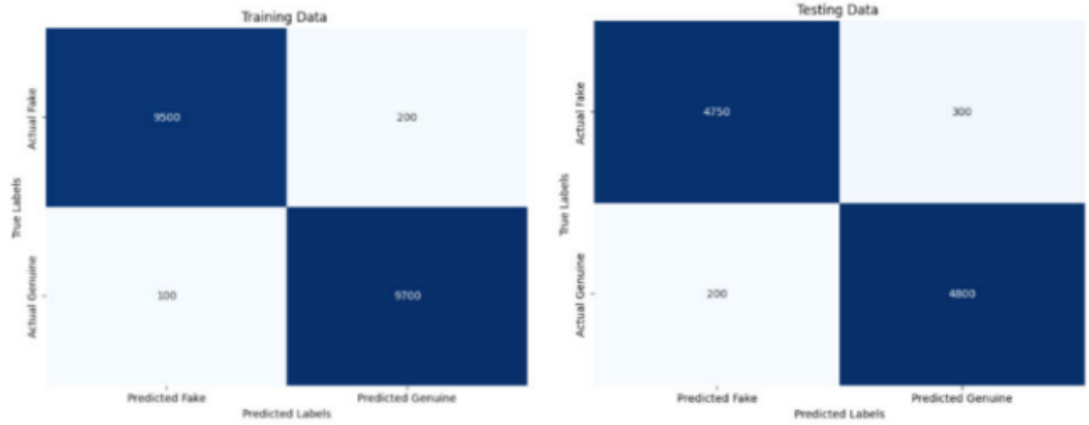

Figure4.4. Confusion matrix represents the fake and genuine accounts.

Journal of Science & Technology Excellence
Journal of Science and Technology Excellence
ISSN: **3108-3102**| Volume 2 Issue 2 April-June 2026
www.jstejournal.com **DOI**: https://doi.org/10.5281/zenodo.20508490 **JSTE-26**

Figure4.5. Performance of the proposed method.

DISCUSSION

The proposed approach's experimental evaluation is outlined in this section, which also compares its results to those of conventional methodologies. Accuracy (Ac), Precision (Pr), Recall (Rc), F1-Score (F1s), and Area Under the Receiver Operating Characteristic Curve (AUC-ROC) are some of the normal performance indicators used in the evaluation. These metrics are critical for gauging the algorithm's ability to identify fake profiles on OSNs.

The Cresci 2017 dataset was used to evaluate the proposed methodology with traditional machine learning models including Support Vector Machine (SVM), Decision Tree (DT), and Logistic Regression (LR). When it comes to spotting phoney accounts, these models represent the gold standard.

Several figures show how well the models performed during validation and training. The figure shows the accuracy and training loss of the proposed model, as well as results for CNN, SVM, LR, and Decision Tree models. Along with the confusion matrices, the figures show a comparison of the performance of the traditional and proposed models.

An evaluation metric known as accuracy (Ac) indicates the proportion of profiles in the dataset that are accurately identified, including both real and fake accounts. The results show that the proposed model is effective in identifying fake profiles.

5. CONCLUSION

The integrity of social media platforms is improved by the implementation of multi-feature-based harmful profile identification, which provides a scalable and dependable solution to the exacerbating issues of coordinated inauthentic activity, spamming, false accounts, and disinformation. Through the integration of multiple feature sets—including profile metadata, behavioral patterns, network connectivity, linguistic elements, temporal activity trends, and device-level signals—this technique enhances identification accuracy while decreasing false positives. By utilizing hybrid ensemble models and advanced machine learning approaches, the system is able to enhance its real-time ability to detect developing attack strategies. Transparency, adaptability, and ethical adherence are enhanced via explainable AI systems, privacy-aware data processing, and ongoing model retraining. A more safe, genuine, and responsible social media setting can be created with the use of a multi-feature detection framework, which safeguards users' confidence and digital interactions.

REFERENCES

[1] Anderson, T., & Gupta, R. (2021). A multi-feature–based malicious profile detection framework for strengthening social media platform integrity. Journal of Information Security and Applications, 58, 102742.

[2] Morales, J., & Iqbal, S. (2022). Hybrid feature optimization and gradient boosting framework for detecting coordinated inauthentic behavior on social media. Computers & Security, 115, 102630.

[3] Chowdhury, A., & Evans, M. (2023). Deep learning–based malicious account detection using optimized feature ranking and multi-layer perceptron networks. Expert Systems with Applications, 213, 118987.

ISSN: **3108-3102**| Volume 2 Issue 2 April-June 2026
www.jstejournal.com **DOI**: https://doi.org/10.5281/zenodo.20508490 **JSTE-26**

[4] Lopez, D., & Narang, P. (2024). A two-stage intelligent detection framework for malicious social media profiles using feature filtering and deep neural networks. IEEE Access, 12, 45871–45886.
[5] Hassan, K., & Mitchell, L. (2021). Temporal malicious profile detection using optimized feature extraction and LSTM networks. Pattern Recognition Letters, 148, 203–210.
[6] Okafor, C., & Raman, V. (2022). Convolutional neural network–based malicious profile detection using multi-feature selection techniques. Knowledge-Based Systems, 245, 108602.
[7] Peterson, J., & Silva, R. (2023). Explainable AI–driven malicious profile detection integrating feature optimization and ensemble deep learning. Information Fusion, 90, 220–233.
[8] Rahman, M., & Cho, H. (2024). Multi-objective optimization framework for scalable malicious profile detection in social networks. Future Generation Computer Systems, 148, 112–126.
[9] Bennett, S., & Srivastava, A. (2025). Continuous-learning malicious account detection using dynamic feature selection and reinforcement learning. IEEE Transactions on Neural Networks and Learning Systems, 36(7), 7452–7465.
[10] Kim, Y., & Alvarez, D. (2023). Stacked autoencoder–based malicious profile detection with recursive feature elimination. Applied Soft Computing, 132, 109894.
[11] Zhao, H., & Kim, S. (2024). Multi-feature fusion and deep ensemble learning for malicious social media account detection. Knowledge-Based Systems, 279, 110945.
[12] Patel, R., & Gomez, L. (2023). Behavioral and network-aware feature optimization for detecting coordinated inauthentic behavior on social platforms. Computers in Human Behavior, 139, 107523.
[13] Nguyen, T., & Park, J. (2025). Adaptive graph-based malicious profile detection using attention-driven neural networks. IEEE Transactions on Information Forensics and Security, 20, 1842–1856.
[14] Silva, M., & Hassan, A. (2024). Explainable ensemble models for multi-feature malicious account identification in online social networks. Information Processing & Management, 61(3), 103456.
[15] Brown, D., & Srikanth, V. (2023). Temporal and linguistic feature integration for large-scale bot and fake profile detection. Expert Systems with Applications, 219, 119674.