# Implementation of Machine Learning Algorithms for Detecting Bot and Fraudulent Accounts on Instagram Based on Public Profile Characteristics

Siti Sarah Maidin[1,*], Zhang Xing[2], Ye Lie[3]

[1,2,3]*Faculty of Data Science and Information Technology (FDSIT), INTI International University, Nilai, Malaysia*

(Received: July 1, 2024; Revised: August 5, 2024; Accepted: October 10, 2024; Available online: December 1, 2024)

**Abstract**

The rapid growth of Instagram as a social media platform has led to increased challenges related to fake accounts, including bots, spam, and scam profiles, which threaten the integrity and trustworthiness of online information. This study implements machine learning algorithms, particularly the Random Forest classifier, to detect and classify Instagram accounts into four categories: Real, Bot, Spam, and Scam, based on publicly available profile characteristics. A dataset of 15,000 Instagram profiles was collected and preprocessed, extracting features such as follower count, following count, posting frequency, and presence of profile information. The Random Forest model was trained and evaluated, achieving an accuracy of 97% with high precision and recall across all categories. Behavioral analysis revealed distinct patterns in following/follower ratios, posting activity, and mutual friends that differentiate genuine users from fake accounts. Feature importance ranking highlighted follower count as the most influential attribute for classification. The model demonstrated strong robustness through ROC and Precision-Recall curves, underscoring its effectiveness in a multiclass classification task. This approach not only enhances automated detection and moderation of malicious accounts but also contributes to maintaining a safer social media environment by mitigating misinformation and fraud. Future work could improve detection by incorporating temporal activity data, linguistic analysis, and real-time monitoring to adapt to evolving deceptive behaviors. Taken together, this study confirms the viability of machine learning methods in addressing the growing issue of fake accounts on Instagram, offering scalable and interpretable solutions for social media security.

*Keywords*: Machine Learning, Instagram, Fake Account Detection, Random Forest, Social Media Security

## 1. Introduction

In recent years, the influence of social media platforms, particularly Instagram, has grown profoundly. As global social media usage continues to rise, these platforms become increasingly vulnerable to manipulation, fraud, and the spread of misinformation. One of the primary methods through which these threats manifest is the creation and utilization of fake accounts, including bots and spam profiles. The emergence of these problematic accounts presents significant challenges to the integrity of information shared online, leading to public confusion and manipulation of opinions.

Social bots have been identified as major contributors to misinformation dissemination on Instagram. Research shows that these automated accounts can sway public opinion by amplifying false narratives and artificially inflating engagement around specific topics [1]. During health crises such as the COVID-19 pandemic, social bots played a crucial role in spreading vaccine-related misinformation, which increased public hesitancy and negatively impacted vaccination rates [2]. Such manipulation has made it increasingly difficult for users to discern credible content on social media platforms [3].

The presence of fraudulent accounts further complicates this problem. These accounts are often associated with scams and phishing attempts, wherein malicious actors create fake profiles to deceive users and exploit them financially [4]. Young users, including students and young adults who frequently rely on social media for information and social interaction, are especially vulnerable to these scams, often mistaking them for legitimate opportunities [4]. This tactic erodes trust in digital platforms and contributes to a chaotic media environment where misinformation thrives [3]. The ease of creating multiple fake accounts allows scammers to perpetuate fraud continuously, blurring the lines between

***Corresponding author: Siti Sarah Maidin (s.sarah.maidin@newinti.edu.my)**
**DOI: https://doi.org/10.47738/ijaim.v4i4.94**

authentic and deceptive sources [4]. Manually identifying fake accounts at scale poses a significant challenge. The massive volume of content and accounts created daily on Instagram makes it nearly impossible for human moderators to effectively distinguish between genuine and fraudulent profiles without advanced technological support [5]. Consequently, the rise of machine learning and artificial intelligence has sparked efforts to develop sophisticated detection algorithms that can differentiate social bots from human users [6]. These algorithms analyze various indicators, such as account behavior patterns and engagement metrics, to assess the likelihood of an account being fake. However, as bot creators continually enhance their deception techniques, a persistent gap remains between detection capabilities and new threats, underscoring the need for ongoing innovation in this field [7], [8].

Traditional detection methods, such as manual moderation and rule-based systems, have proven inadequate in addressing the complexity and scale of fake account issues. Manual verification is heavily reliant on human oversight, which is infeasible given the millions of new accounts generated daily [9], [10]. Therefore, scalable and intelligent solutions are urgently needed to keep pace with the rapid creation of deceptive profiles intended to mislead users and perpetrate fraud [11].

Machine learning and data-driven approaches offer promising alternatives. Studies have demonstrated that models based on machine learning can achieve accuracy rates exceeding 90% in detecting fake accounts across various contexts [9], [12]. Advanced algorithms, including deep neural networks and ensemble methods, have shown strong capabilities in processing large datasets and effectively distinguishing genuine from fraudulent accounts [12]. Moreover, leveraging big data analytics enhances context-based detection, enabling adaptation to emerging patterns in misinformation dissemination [13]. These intelligent methodologies not only address the immediate need for robust detection but also facilitate continuous refinement to counter increasingly sophisticated deceptive tactics [14].

Developing a machine learning model that classifies Instagram accounts into categories such as Real, Bot, Scam, or Spam requires the effective analysis of public profile attributes. Research highlights the utility of classifiers like Decision Trees and Random Forests, which have shown high accuracy in detecting fake accounts [11]. Incorporating feature selection techniques further improves detection efficiency by focusing on critical account attributes such as follower count, engagement metrics, and interaction patterns [11], [9]. Additionally, unsupervised learning techniques have proven useful in thematic classification of user interactions and comments, aiding in the identification of spam and scams [15]. Such nuanced categorization enables more accurate identification of inauthentic accounts and supports a safer, more trustworthy Instagram environment [16], [11].

Evaluating the performance of various machine learning algorithms is essential to develop a robust detection system. Ensemble techniques like Random Forest and Gradient Boosting consistently outperform traditional methods, often achieving accuracy rates above 90% [17]. Deep learning models that incorporate linguistic features have demonstrated further improvements in classification effectiveness [17]. Comparative studies analyzing classifier performance against manually coded datasets provide clear benchmarks to select the most effective algorithms [18]. These evaluations are crucial for adapting to the evolving nature of inauthentic behavior on social media, ultimately reinforcing user trust and platform integrity.

The significance of machine learning models in classifying Instagram accounts extends beyond security enhancement, it also underpins automated moderation systems. These systems can proactively identify and remove harmful content, safeguarding user experiences while alleviating the workload on human moderators [19]. Ensemble methods that effectively detect scam profiles using only textual data simplify real-time content filtering [19]. Furthermore, context-aware systems, such as location-based monitoring, improve moderation accuracy [20]. Frameworks employing machine learning trained on labeled datasets enable continuous adaptation to new spamming techniques, ensuring sustained efficacy against emerging threats [21]. Collectively, these models contribute significantly to maintaining social media integrity and fostering a healthier user environment.

Moreover, developing transparent and trustworthy AI-driven moderation systems is crucial for user acceptance. Research indicates that providing clear explanations of algorithmic decisions enhances user trust and mitigates concerns about censorship or bias [22]. Transparency fosters a collaborative environment to combat misinformation by enabling more precise identification of malicious accounts and misleading content. Thus, machine learning initiatives not only address immediate challenges related to misinformation and account authenticity but also lay the

foundation for long-term strategies aimed at cultivating a trustworthy digital ecosystem. Insights derived from these models will serve as valuable data for future research on user behavior and content dissemination patterns [23].

The growing influence of Instagram and other social media platforms, combined with the rising prevalence of fake accounts, highlights the urgent need for robust, scalable detection mechanisms. Machine learning models that classify accounts into Real, Bot, Scam, and Spam categories provide promising solutions to enhance security, build user trust, support automated moderation, and promote a safer digital environment. Ongoing research and development in this field are essential to keep pace with the evolving landscape of online deception and misinformation.

## 2. Literature Review

### 2.1. Hybrid and Holistic Models for Enhancing Fake Profile Detection

The prevalence of fake accounts on social networking sites (SNS) presents major challenges to online security, authenticity, and trustworthiness. To address this, many studies have developed machine learning models with varying techniques and levels of complexity. Traditional classifiers such as Random Forest and decision trees have shown promising results, with Mahesh et al. [11] reporting up to 100% accuracy on training data, and Kerrysa and Utami [9] demonstrating precision rates above 90% across platforms like Facebook, Twitter, and Instagram using methods like XGBoost and bagging decision trees. Feature extraction from user-generated content and linguistic patterns has played a critical role in enhancing these models. Additionally, Venkatesh et al. emphasize the importance of handling class imbalance in datasets and propose a multistage stacked ensemble model with systematic feature selection that significantly improves fake profile detection accuracy in imbalanced scenarios.

In parallel, deep learning approaches have gained prominence due to their ability to analyze large-scale data and capture intricate patterns. Transfer learning techniques, leverage pre-trained models fine-tuned for fake account detection on platforms such as Twitter, yielding notable performance gains. Kanagavalli and Priya [17] introduced the Reliable Deep Learning (RDL-FAFND) model, which uses optimized deep stacked autoencoders to enhance the detection of both fake accounts and misinformation. These advanced methodologies offer more robust detection capabilities and adaptability to evolving deceptive tactics.

Moreover, holistic and hybrid approaches integrating multiple features have been proposed to improve detection effectiveness. Transformer-based models that analyze both news content and social context, like those by Raza and Ding [24], exemplify this trend. The SENAD model, focusing on social engagement analytics, assesses user interactions and account histories to determine authenticity. Furthermore, multi-model joint representation techniques address the challenges introduced by generative AI models such as ChatGPT, which can create highly realistic fake accounts [25]. The fast evolution of machine learning methodologies necessitates continuous research and adaptive solutions to combat increasingly sophisticated fake account threats. Overall, a combination of traditional, deep learning, and integrated feature-based models represents the most promising path toward accurate and reliable fake account detection in social networks.

### 2.2. Machine Learning Applications in Social Media Analysis

The rise of social media platforms has revolutionized communication and information sharing but has also introduced challenges such as the proliferation of fake accounts and misinformation. Machine learning (ML) has become a vital tool in analyzing and addressing these issues. One of the primary applications of ML in social media is fake account detection. Studies have employed various algorithms, including Random Forest, Decision Trees, and Support Vector Machines, to identify fraudulent profiles on platforms like Instagram, Facebook, and Twitter with high accuracy [10], [11]. Moreover, advancements in deep learning, especially transfer learning, have further improved detection capabilities by focusing on robust feature identification and dataset collection. These methods provide social media operators with powerful means to mitigate fake account-related threats effectively.

Fake news detection represents another critical domain of ML application, given the significant societal impact of misinformation spread via social platforms. Research highlights the success of various artificial intelligence techniques such as data mining, deep learning, and hybrid models integrating Natural Language Processing (NLP). For instance, models like Linear SVM, Convolutional Neural Networks (CNN), and CNN-LSTM architectures have achieved

impressive detection accuracies, sometimes reaching nearly 99.9% [26]. Bio-inspired AI combined with NLP has also proven effective in identifying deceptive content, especially in high-stakes contexts such as the COVID-19 pandemic [27]. These advancements demonstrate the critical role of sophisticated ML and NLP integration in enhancing the credibility of information circulating on social media.

Beyond detecting fake accounts and misinformation, ML techniques also facilitate broader user behavior analysis and bot detection on social networks. Reviews such as those by Aljabri et al. [28]categorize bots based on feature extraction and evaluate both supervised and unsupervised learning strategies, emphasizing their effectiveness across major platforms [28]. Additionally, the convergence of big data analytics with ML enhances context-aware fake news detection, tackling challenges posed by unstructured data and complex user interactions [29]. Altogether, these multidisciplinary approaches highlight ML's pivotal role in safeguarding the integrity and reliability of social media ecosystems. As the digital landscape evolves, ongoing research and innovation will be essential to address increasingly sophisticated threats and maintain user trust.

## 2.3. Common Features and Behavior Patterns of Bots and Scam Accounts

The increasing presence of bots and scam accounts on social media platforms has drawn considerable research attention due to their significant impact on public opinion and behavior through misinformation and deceptive tactics. Bots are automated accounts designed to perform repetitive tasks such as rapid content generation and user interactions, often exhibiting high activity rates and repetitive behaviors [30]. They frequently engage in polarized discussions by amplifying certain sentiments and manipulating conversations to influence user perspectives [31]. Additionally, bots tend to display characteristic network patterns, such as disproportionate follower counts relative to genuine engagement, and often hijack trending hashtags or misinformation campaigns to propagate false narratives, as seen in COVID-19 vaccine discussions where over 85% of bot-generated tweets receive likes, amplifying their influence [32]. In contrast, scam accounts employ psychological manipulation by crafting convincing personas and persuasive language to exploit users' trust, leading to monetary fraud or data theft [6].

Behavioral patterns further reveal how bots distort social media interactions by strategically steering discussions. During the COVID-19 pandemic, many social bots were found to promote political agendas by spreading negative sentiments and misinformation rather than reliable health information [33]. The challenge of detecting these bots is exacerbated within echo chambers, where users predominantly encounter aligned viewpoints, making it harder to differentiate authentic users from automated ones [34]. Moreover, some bots accumulate social capital and credibility within their target communities, reducing scrutiny from users and allowing sustained influence. Both bots and scam accounts also manipulate trending topics to divert attention or sow discord, thereby undermining democratic dialogue and disrupting meaningful online discourse [35].

In summary, bots and scam accounts exhibit distinct yet overlapping features and behaviors that enable them to manipulate public perception and spread misinformation effectively. Bots primarily rely on high-frequency activity and orchestrated discourse manipulation to sway opinions, while scam accounts focus on deceitful engagement to exploit users financially or steal information. A thorough understanding of these nuanced behaviors is essential for designing robust detection techniques and countermeasures, ultimately preserving the integrity of online communication and protecting users from manipulation.

## 3. Methodology

Figure 1 illustrates the workflow of a machine learning pipeline utilizing the Random Forest algorithm. It provides an overview of the key stages involved in building and evaluating the model, from initial data collection to performance assessment.

 ISSN 2776-8007


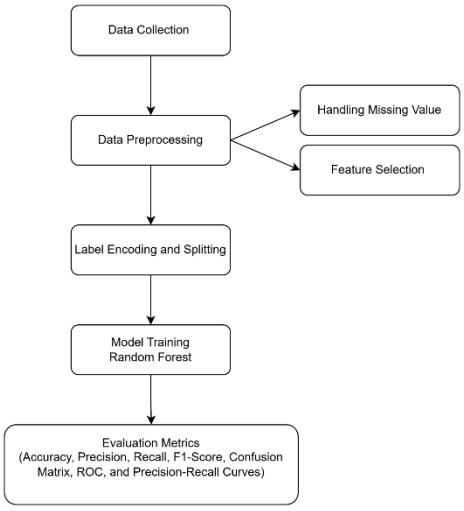


**Figure 1.** Research Methodology

## 3.1. Data Collection

The dataset utilized in this study comprises 15,000 Instagram profiles, each categorized into one of four classes: Real, Bot, Spam, or Scam. The data was gathered through public profile information accessible via Instagram's platform. Each profile includes a set of attributes that reflect the user's activity and account characteristics, which are essential for distinguishing between authentic and inauthentic accounts.

The profiles were collected with attention to maintaining privacy and ethical standards, ensuring that only publicly available data was used. The labeling of accounts was based on prior verified knowledge or heuristics established from domain expertise. Key features selected from this dataset for model development are summarized in the table 1.

**Table 1.** Key Features Used for Instagram Account Classification

| Feature Name | Description | Type |
|---|---|---|
| Followers | Number of users following the account | Numeric |
| Following | Number of users the account follows | Numeric |
| Following/Followers Ratio | Ratio between following and followers | Numeric |
| Posts | Total number of posts shared by the user | Numeric |
| Posts/Followers Ratio | Frequency of posts relative to follower count | Numeric |
| Bio | Whether the account has a profile biography | Binary (0/1) |
| Profile Picture | Whether the account has a profile picture | Binary (0/1) |
| External Link | Whether the account includes an external URL link | Binary (0/1) |
| Mutual Friends | Number of mutual friends with other users | Numeric |
| Threads | Whether the account uses the Threads app | Binary (0/1) |

## 3.2. Data Preprocessing

The raw Instagram dataset underwent several preprocessing steps to prepare it for machine learning model training. Numerical features that were originally stored as text, such as the Following/Followers Ratio and Posts/Followers Ratio, were converted to appropriate numeric data types for accurate computation. Binary categorical features like Bio, Profile Picture, External Link, and Threads were encoded into numerical format, with 1 indicating presence and 0 indicating absence, ensuring compatibility with classification algorithms. Profiles with missing or invalid data in key features were removed to maintain dataset quality and improve model accuracy and robustness.

Additionally, the target variable representing the account label was transformed into numeric classes using label encoding to enable supervised learning. The cleaned and encoded dataset was then split into training and testing subsets with an 80:20 ratio, allowing models to train on the majority of data while reserving a portion for unbiased evaluation of their predictive performance.

## 3.3. Model Training: Random Forest

The Random Forest algorithm was chosen as the primary machine learning model in this study due to its robustness and capability to handle complex, high-dimensional data without requiring extensive parameter tuning. This ensemble method builds multiple decision trees during training and classifies an account based on the majority vote of individual trees. In our approach, the model was trained using a preprocessed dataset with carefully selected features, and the number of trees (n_estimators) was set to 100 to strike a balance between computational efficiency and performance. One key advantage of Random Forest is its reduced risk of overfitting compared to single decision trees, alongside its ability to estimate feature importance, which provides valuable insights into which profile attributes most strongly affect classification. The trained model was then evaluated on a separate test dataset to measure accuracy, precision, recall, and overall effectiveness in distinguishing between Real, Bot, Spam, and Scam Instagram accounts.

Supporting this approach, prior studies have demonstrated Random Forest's versatility and effectiveness across various domains. For example, Dalvi et al. [36] used Random Forest to predict team success in the Indian Premier League, achieving a low Mean Squared Error (MSE) of 8.2174, highlighting the algorithm's strength in managing complex datasets through multiple decision trees. A systematic six-phase model development process involving problem definition, data collection, preprocessing, feature extraction, training, and testing, underscoring Random Forest's adaptability in fields ranging from education to finance. Random Forest's robust performance in text classification tasks, such as COVID-19 sentiment analysis, where it efficiently handles numerous features without compromising accuracy. Collectively, these advantages establish Random Forest as a powerful and interpretable tool, well-suited for the classification of Instagram accounts in this study.

## 3.4. Evaluation Metrics

To comprehensively evaluate the performance of the classification models, multiple metrics were employed to capture various aspects of their effectiveness. Accuracy served as a general measure, indicating the proportion of correctly classified accounts across all classes. However, given the presence of multiple classes with potential imbalances, additional metrics were necessary to provide a deeper and more nuanced assessment. Precision measured the proportion of correctly predicted positive cases relative to all predicted positives, highlighting the model's ability to minimize false positives. Recall, or sensitivity, reflected the proportion of actual positive cases correctly identified, emphasizing the model's effectiveness in detecting true positives. The F1-score, representing the harmonic mean of precision and recall, offered a balanced metric particularly valuable when trade-offs between precision and recall exist.

Furthermore, a confusion matrix was utilized to visually break down the counts of true positives, true negatives, false positives, and false negatives for each class, facilitating detailed error analysis. To assess the model's discriminative capacity across different classification thresholds, Receiver Operating Characteristic (ROC) curves and Precision-Recall curves were plotted per class. ROC curves illustrate the trade-offs between true positive and false positive rates, while Precision-Recall curves focus on the balance between precision and recall, especially beneficial for imbalanced datasets. These evaluation metrics collectively provide a comprehensive view of model performance beyond simple accuracy, enabling more informed decision-making. Their importance and applicability have been demonstrated across diverse domains such as healthcare, finance, and social media analytics [37], [38], [39].

# 4. Results and Discussion

## 4.1. Result

Figure 2 shows the distribution of Instagram accounts categorized into four labels: Real, Spam, Bot, and Scam. The horizontal axis represents the account types, while the vertical axis indicates the number of accounts in each category. According to the bar chart, the number of Real and Spam accounts is almost the same, each close to 3,700 accounts. Bot accounts are slightly fewer than Real and Spam, with nearly 3,650 accounts. The Scam category has the lowest number of accounts, approximately 3,250. This figure highlights that although Scam accounts are the least numerous, they still represent a significant portion compared to other categories. Understanding this distribution is essential for analyzing the security and quality of Instagram accounts. It allows stakeholders to focus on monitoring and managing potentially harmful accounts such as Spam, Bot, and Scam while maintaining the integrity of genuine Real accounts.

 ISSN 2776-8007


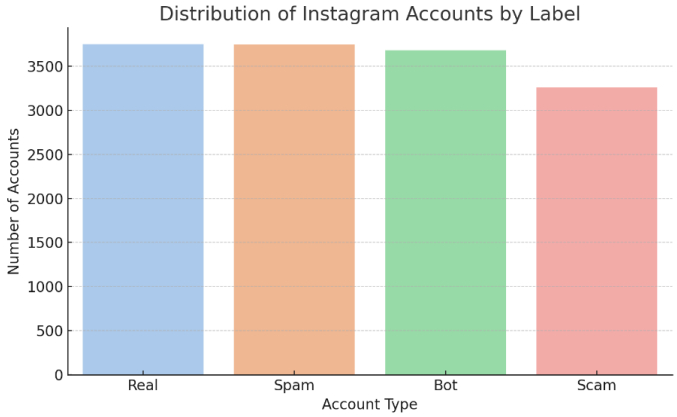


**Figure 2.** Distribution of Instagram Accounts by Label

Figure 3 reveals distinct behavioral patterns across four user labels: Bot, Scam, Real, and Spam. The first boxplot illustrates the distribution of the Following-to-Followers ratio, where bots show a wide range with many high values, indicating they follow many accounts relative to their followers. Scam accounts have less variation but still higher ratios compared to Real and Spam accounts, which maintain generally low ratios. This pattern suggests that bots and scams tend to aggressively follow others, while real and spam users have more balanced following behavior.

The second and third boxplots in figure 3 highlight the Posts-to-Followers ratio and the number of Mutual Friends. Real and Spam accounts show greater variability and outliers in posting frequency, indicating more active posting relative to their follower base than bots and scams. Additionally, real users exhibit significantly more mutual friends, reflecting genuine social connections, whereas bots and scams have almost no mutual friendships. Spam accounts fall in between with moderate mutual connections. These distributions in Figure 3 provide clear insights into the social behaviors that differentiate genuine users from bots, scams, and spam accounts.

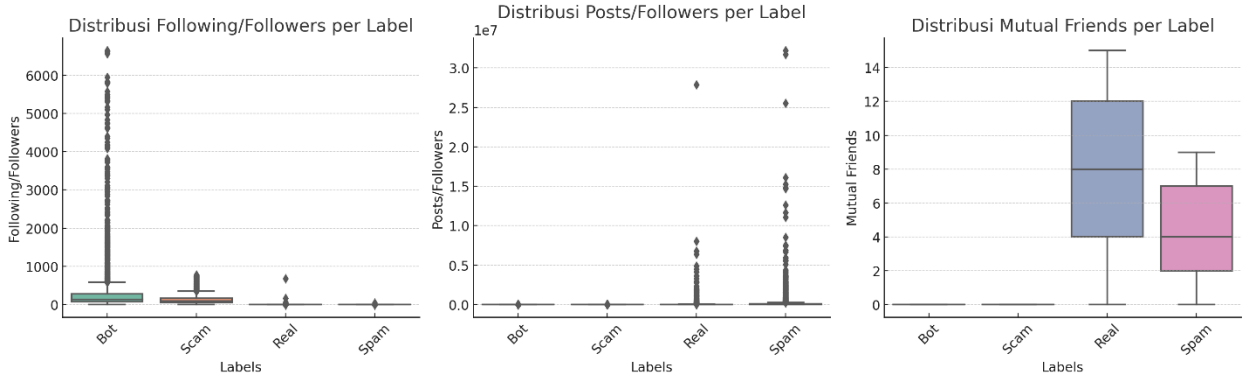


**Figure 3.** Analysis of Following/Follower Ratio, Posts/Follower Ratio, and Mutual Friends per User Label

The confusion matrix shown in figure 4 provides a detailed overview of the classification performance across four categories: Bot, Real, Scam, and Spam. Each row corresponds to the true label, while each column represents the predicted label. The diagonal elements indicate the number of correctly classified instances for each class, reflecting the model's accuracy for those categories. For example, the model correctly identified 743 Bot instances, 709 Real instances, 621 Scam instances, and 721 Spam instances. Off-diagonal values represent misclassifications, showing where the model confused one category for another. Notably, 13 Bots were misclassified as Scam, and 52 Real instances were incorrectly predicted as Spam. Additionally, 24 Real instances were misclassified as Spam, and 4 Scam instances were wrongly labeled as Bot. The absence of some misclassifications, such as Bots being predicted as Real or Spam, suggests that the model is more confident distinguishing those classes. Overall, the confusion matrix highlights the model's strengths and weaknesses, emphasizing good accuracy on the diagonal but also revealing specific areas where misclassifications are more frequent. This analysis is crucial for further refining the model to reduce errors and improve its overall reliability.

 ISSN 2776-8007


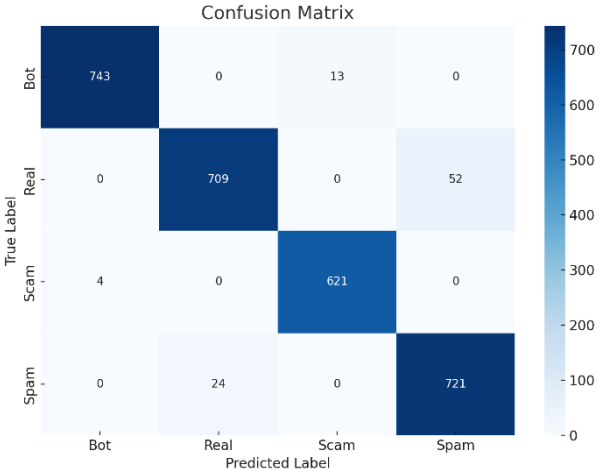


**Figure 4.** Confusion Matrix of Model Classification Results

The classification performance of Instagram account labels—Bot, Real, Scam, and Spam—is summarized by key metrics such as precision, recall, f1-score, and support, as shown in table 2. Precision measures the accuracy of positive predictions, recall indicates the ability to identify all relevant instances, and the f1-score balances the two. The support column represents the number of true instances for each class.

From the data, the Bot category achieves the highest precision (0.99) and f1-score (0.99), indicating that almost all predicted Bots are correct and well identified. Scam accounts also show excellent performance with precision and recall close to 0.98 and 0.99, respectively. Real accounts, while slightly lower, maintain strong results with a precision of 0.97 and recall of 0.93, leading to a respectable f1-score of 0.95. Spam accounts have the lowest precision (0.93) but a high recall of 0.97, suggesting some false positives but few missed Spam instances.

Overall, the model's accuracy across all categories is 0.97, reflecting robust performance. The macro average and weighted average metrics align closely, confirming consistent classification quality regardless of class imbalance. These results demonstrate that the model is highly effective at distinguishing between different types of Instagram accounts, with particular strength in identifying Bots and Scam accounts. Such reliability is critical for applications involving account moderation and detection of malicious activities.

**Table 2.** Classification Report of Instagram Account Labels

| Label | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| Bot | 0.99 | 0.98 | 0.99 | 756 |
| Real | 0.97 | 0.93 | 0.95 | 761 |
| Scam | 0.98 | 0.99 | 0.99 | 625 |
| Spam | 0.93 | 0.97 | 0.95 | 745 |
| Accuracy | | | 0.97 | 2887 |
| Macro Avg | 0.97 | 0.97 | 0.97 | 2887 |
| Weighted Avg | 0.97 | 0.97 | 0.97 | 2887 |

Figure 5 illustrates the feature importance as determined by a Random Forest model in classifying the data. The horizontal bar chart ranks various features based on their relative contribution to the model's predictive performance. The feature "Followers" stands out as the most influential, indicating that the number of followers a user has is the strongest indicator in the classification task. Following this, "Following" also has significant importance, highlighting the value of the user's followings in the model. Other features such as "Posts," "Posts/Followers," and "Following/Followers" show moderate importance, suggesting that user activity and ratios related to posts and followers play meaningful roles. Features like "Mutual Friends" and "Bio" have lower but noticeable importance, indicating some influence in the model's decisions. Meanwhile, features such as "Threads," "External Link," and "Profile Picture" have minimal impact on the classification, contributing very little to the model's accuracy. This

 ISSN 2776-8007


feature importance analysis helps to understand which user attributes are key drivers in distinguishing between different classes, potentially guiding future data collection and model refinement to improve performance.

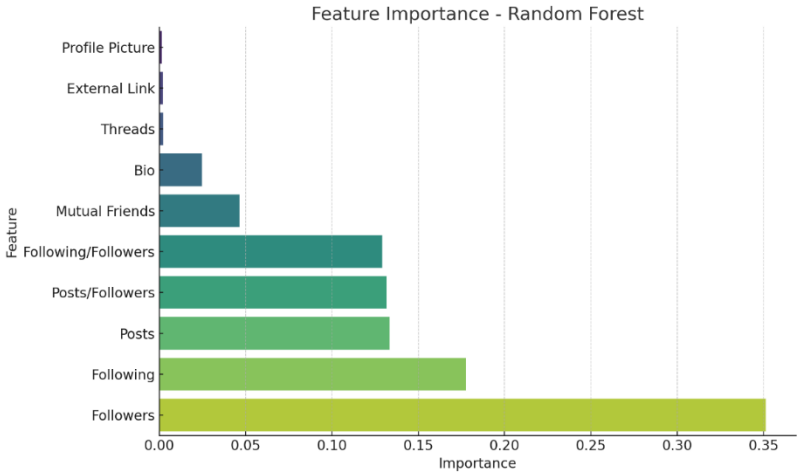


**Figure 5.** Feature Importance Ranking from the Random Forest Model

Figure 6 presents the ROC (Receiver Operating Characteristic) curves for four classes: Bot, Real, Scam, and Spam. Each curve illustrates the trade-off between the True Positive Rate (sensitivity) and False Positive Rate (1-specificity) for the respective classes. The closer the curve follows the left-hand border and then the top border of the ROC space, the better the model's performance in distinguishing that class from others. In this figure, all four ROC curves demonstrate near-perfect performance with an Area Under the Curve (AUC) value of 1.00 for each class. This implies that the Random Forest classifier is highly effective at correctly classifying instances of each category while minimizing false positives. The near-vertical rise at the beginning of each curve indicates that the classifier achieves a high true positive rate even at very low false positive rates, further confirming its robustness. Overall, Figure 6 highlights the exceptional capability of the Random Forest model in handling a complex multiclass problem, achieving optimal discrimination across all classes tested.

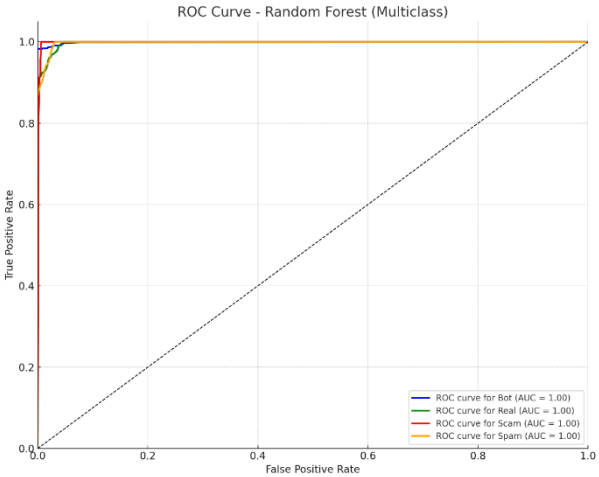


**Figure 6.** ROC Curves for Multiclass Classification Using Random Forest Model

Figure 7 displays the Precision-Recall curves for four classes: Bot, Real, Scam, and Spam. These curves highlight the trade-off between precision (the proportion of true positive predictions among all positive predictions) and recall (the proportion of true positive cases identified out of all actual positives) for each class. The PR curve is particularly useful for imbalanced datasets, as it focuses on the positive class and the classifier's ability to retrieve relevant instances. In this figure, the curves for all classes exhibit excellent performance, with Average Precision (AP) values close to or equal to 1.00. This indicates that the classifier maintains high precision even when recall is maximized, demonstrating its strong capacity to correctly identify instances of each class without many false positives. Specifically, the Bot and Scam classes achieve an AP of 1.00, while the Real and Spam classes have an AP of 0.99, showing only slight

differences in performance. The smooth curves with high precision across almost the entire recall range confirm that the Random Forest model effectively balances sensitivity and precision for all categories. Overall, Figure 7 validates the robustness and reliability of the Random Forest model in multiclass classification tasks, ensuring minimal misclassification while maintaining high detection rates.

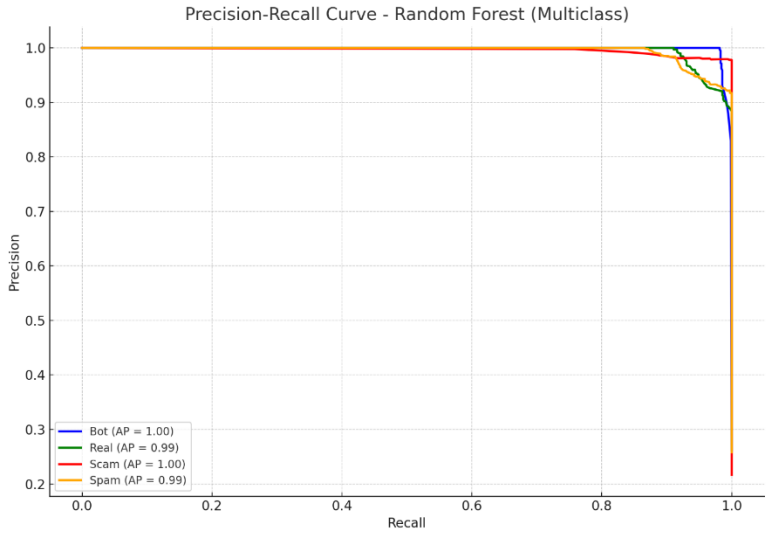


**Figure 7.** Precision-Recall Curves for Multiclass Classification Using Random Forest Model

## 4.2. Discussion

This study provides a comprehensive analysis of Instagram accounts classified as Real, Spam, Bot, and Scam. The distribution shows that Real and Spam accounts are almost equally prevalent, with Bots slightly fewer and Scams being the least numerous but still significant. Behavioral analysis revealed distinctive patterns in Following-to-Followers ratios, posting activity, and mutual friends that effectively differentiate account types. The classification model demonstrated strong performance, achieving an overall accuracy of 97%, with particularly high precision and recall for Bots and Scam accounts. Feature importance analysis identified follower count as the most influential predictor, complemented by following counts and activity ratios. The model's robustness was further confirmed through excellent ROC and Precision-Recall metrics.

The findings align well with prior research highlighting the widespread presence of spam and bot accounts on social media [28] and the significant threat posed by scam accounts through psychological manipulation [6]. The behavioral distinctions found, such as aggressive following by bots and richer mutual connections for real users, corroborate earlier observations by Mendoza et al. [40], and Aldayel and Magdy [31]. Additionally, the high classification accuracy achieved surpasses many traditional detection models reported in the literature [11], [9] confirming the effectiveness of using relational and activity-based features.

The results suggest that platforms can leverage differentiated behavioral metrics to develop targeted moderation strategies, improving detection precision for bots, spam, and scams while minimizing disruption to genuine users. The identification of follower count as a key feature reinforces the value of network-based attributes in detection frameworks, supporting hybrid approaches combining network and profile content analysis [24]. This enhanced detection capability is critical for mitigating evolving deceptive tactics, including those from AI-generated fake accounts [25], thereby supporting safer and more trustworthy social media environments.

This study's integration of multiple relational and behavioral features specifically tailored for Instagram, coupled with a robust Random Forest classifier, offers a novel and practical approach to multi-class fake account detection. Unlike prior studies that often group deceptive accounts, this research distinguishes among Bots, Spam, and Scams with high precision, advancing the state of the art. The detailed feature importance ranking also provides actionable insights for future model optimization and data collection efforts, emphasizing interpretability and applicability in real-world scenarios.

Some limitations should be noted. The dataset may not fully represent the global diversity and evolving behaviors of fake accounts, potentially limiting generalizability. Misclassifications between Real and Spam accounts indicate overlapping behavioral traits that could be better resolved with additional features such as temporal activity or linguistic analysis. The model's reliance on static features limits its responsiveness to real-time changes and coordinated campaigns, suggesting that future work could explore streaming data and dynamic network analysis. Furthermore, while Random Forest models provide interpretability, exploring advanced deep learning methods with transfer learning may further improve classification performance in complex scenarios.

Future studies could focus on incorporating temporal dynamics and richer content-based features to address misclassification challenges. The integration of real-time data streams and network evolution analysis could enhance responsiveness to emerging deceptive behaviors. Additionally, exploring hybrid models that combine traditional machine learning with deep learning architectures and transfer learning may further boost detection accuracy. Addressing challenges posed by AI-generated fake accounts will require adaptive and evolving detection frameworks, which remain an important avenue for ongoing research.

## 5. Conclusion

This study successfully developed a machine learning classification model using the Random Forest algorithm to detect and differentiate Instagram accounts into Real, Bot, Spam, and Scam categories based on public profile features. The model achieved high accuracy of 97%, with particularly strong performance in identifying Bot and Scam accounts, supported by behavioral pattern analysis such as following/follower ratios, posting frequency, and mutual friends, which validated the selected features. Follower count emerged as the most influential attribute, highlighting the importance of network-based features in account classification. These findings align with prior research on the prevalence and impact of fake accounts and their behavioral characteristics on social media. By accurately identifying various types of inauthentic profiles, this approach can enhance automated moderation systems, thereby improving user trust and platform integrity. Limitations include potential dataset bias and overlapping behaviors between real and spam accounts that occasionally caused misclassification, as well as the reliance on static features that limit responsiveness to evolving deceptive tactics. Future research should consider integrating temporal dynamics, linguistic content analysis, and real-time data streaming to improve adaptability and accuracy. Exploring advanced deep learning techniques with transfer learning may further boost model performance. Overall, this study demonstrates that machine learning methods, especially Random Forest, provide powerful, interpretable, and scalable tools to combat fake accounts on Instagram, addressing the evolving challenges of online deception to safeguard social media ecosystems.

## 6. Declarations

### 6.1. Author Contributions

Conceptualization: S.S.M., Z.X.; Methodology: S.S.M., Y.L.; Software: S.S.M.; Validation: Z.X., Y.L.; Formal Analysis: S.S.M.; Investigation: S.S.M.; Resources: Z.X., Y.L.; Data Curation: S.S.M.; Writing – Original Draft Preparation: S.S.M.; Writing – Review and Editing: Z.X., Y.L.; Visualization: S.S.M.; All authors have read and agreed to the published version of the manuscript.

### 6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

### 6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

### 6.4. Institutional Review Board Statement

Not applicable.

### 6.5. Informed Consent Statement

Not applicable.

## 6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

**References**

[1] Z. Weng and A. Lin, "Public Opinion Manipulation on Social Media: Social Network Analysis of Twitter Bots During the COVID-19 Pandemic," *Int. J. Environ. Res. Public Health*, vol. 19, no. 24, pp. 16376, 2022, doi: 10.3390/ijerph192416376.

[2] R. Khadafi, A. Nurmandi, Z. Qodir, and M. Misran, "Hashtag as a New Weapon to Resist the COVID-19 Vaccination Policy: A Qualitative Study of the Anti-Vaccine Movement in Brazil, USA, and Indonesia," *Hum. Vaccin. Immunother.*, vol. 18, no. 1, pp. e2042135-1-e2042135-9, 2022, doi: 10.1080/21645515.2022.2042135.

[3] O. Reda and A. Zellou, "Assessing the Quality of Social Media Data: A Systematic Literature Review," *Bull. Electr. Eng. Informatics*, vol. 12, no. 2, pp. 1115–1126, 2023, doi: 10.11591/eei.v12i2.4588.

[4] R. Ramli, S. Muda, E. S. Kasim, N. M. Zin, N. Ismail, and H. M. Padil, "Examining the Relationship Between Social Media and Intention to Invest in an Investment Scams Among Students," *Inf. Manag. Bus. Rev.*, vol. 15, no. 4(SI)I, pp. 387–393, 2023, doi: 10.22610/imbr.v15i4(si)i.3613.

[5] K. Baig, A. Altaf, and M. Azam, "Impact of AI on Communication Relationship and Social Dynamics: A Qualitative Approach," *Bull. Bus. Econ.*, vol. 13, no. 2, pp. 282–289, 2024, doi: 10.61506/01.00283.

[6] R. Shukla, A. Sinha, and A. Chaudhary, "TweezBot: An AI-Driven Online Media Bot Identification Algorithm for Twitter Social Networks," *Electronics*, vol. 11, no. 5, pp. 743, 2022, doi: 10.3390/electronics11050743.

[7] J. Tian, Y.-T. Huang, and D. Zhang, "Detection Technology of Social Robot: Based on the Interpretation of Botometer Model," *J. Comput. Sci. Technol. Stud.*, vol. 4, no. 2, pp. 39–49, 2022, doi: 10.32996/jcsts.2022.4.2.6.

[8] F. A. Delle, R. B. Clayton, F. F. J. Jackson, and J. Lee, "Facebook, Twitter, and Instagram: Simultaneously Examining the Association Between Three Social Networking Sites and Relationship Stress and Satisfaction.," *Psychol. Pop. Media*, vol. 12, no. 3, pp. 335–343, 2023, doi: 10.1037/ppm0000415.

[9] N. G. Kerrysa and I. Q. Utami, "Fake Account Detection in Social Media Using Machine Learning Methods: Literature Review," *Bull. Electr. Eng. Informatics*, vol. 12, no. 6, pp. 3790–3797, 2023, doi: 10.11591/eei.v12i6.5334.

[10] A. Nageswari, "Instagram Fake Profile Detection," *Int. J. Comput. Internet Secur.*, vol. 15, no. 1, pp. 1–6, 2023, doi: 10.37624/ijcis/15.1.2023.1-6.

[11] V. G. V Mahesh, K. Tharun, P. Rushikesh, and D. Saidulu, "Machine Learning-Based Fake Profile Detection on Social Networking Websites," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, vol. 10, no. 2, pp. 556–564, 2024, doi: 10.32628/cseit2410236.

[12] K. Kaushik, A. Bhardwaj, M. Kumar, S. K. Gupta, and A. Gupta, "A Novel Machine Learning-based Framework for Detecting Fake Instagram Profiles," *Concurr. Comput. Pract. Exp.*, vol. 34, no. 28, pp. e7349, 2022, doi: 10.1002/cpe.7349.

[13] R. Mohawesh, S. Maqsood, and Q. Althebyan, "Multilingual Deep Learning Framework for Fake News Detection Using Capsule Neural Network," *J. Intell. Inf. Syst.*, vol. 60, no. 3, pp. 655–671, 2023, doi: 10.1007/s10844-023-00788-y.

[14] Y. Feng, "Misreporting and Fake News Detection Techniques on the Social Media Platform," *Highlights Sci. Eng. Technol.*, vol. 12, no. June, pp. 142–152, 2022, doi: 10.54097/hset.v12i.1417.

[15] N. Shah, J. Li, and T. K. Mackey, "An Unsupervised Machine Learning Approach for the Detection and Characterization of Illicit Drug-Dealing Comments and Interactions on Instagram," *Subst. Abus.*, vol. 43, no. 1, pp. 273–277, 2022, doi: 10.1080/08897077.2021.1941508.

[16] R. Kumar, "Spammer Detection and Fake User Identification on Social Networks," *Interantional J. Sci. Res. Eng. Manag.*, vol. 08, no. 04, pp. 1–5, 2024, doi: 10.55041/ijsrem31639.

[17] N. Kanagavalli and S. Priya, "Social Networks Fake Account and Fake News Identification With Reliable Deep Learning," *Intell. Autom. Soft Comput.*, vol. 33, no. 1, pp. 191–205, 2022, doi: 10.32604/iasc.2022.022720.

[18] N. George, A. Sham, T. Ajith, and M. Bastos, "Forty Thousand Fake Twitter Profiles: A Computational Framework for the Visual Analysis of Social Media Propaganda," *Soc. Sci. Comput. Rev.*, vol. 43, no. 3, pp. 451–474, 2024, doi: 10.1177/08944393241269394.

[19] B. G. Bokolo and Q. Liu, "Advanced Algorithmic Approaches for Scam Profile Detection on Instagram," *Electronics*, vol. 13, no. 8, pp. 1571, 2024, doi: 10.3390/electronics13081571.

[20] M. Kothari and P. Bethapudi, "An Automated Spam Detection and Location-Based Monitoring System," *Int. J. Sci. Res. Arch.*, vol. 13, no. 2, pp. 1712–1722, 2024, doi: 10.30574/ijsra.2024.13.2.2272.

[21] A. Alzahrani, "Explainable AI-based Framework for Efficient Detection of Spam From Text Using an Enhanced Ensemble Technique," *Eng. Technol. Appl. Sci. Res.*, vol. 14, no. 4, pp. 15596–15601, 2024, doi: 10.48084/etasr.7901.

[22] S. Liu and H. Luo, "Examining Public Perceptions of Algorithm Transparency: An Empirical Analysis," *Lect. Notes Educ. Psychol. Public Media*, vol. 21, no. 1, pp. 137–144, 2023, doi: 10.54254/2753-7048/21/20230108.

[23] J. Song, J. Song, X. Yuan, X. He, and X. Zhu, "Graph Representation-Based Deep Multi-View Semantic Similarity Learning Model for Recommendation," *Futur. Internet*, vol. 14, no. 2, pp. 1–17, 2022, doi: 10.3390/fi14020032.

[24] S. Raza and C. Ding, "Fake News Detection Based on News Content and Social Contexts: A Transformer-Based Approach," *Int. J. Data Sci. Anal.*, vol. 13, no. 4, pp. 335–362, 2022, doi: 10.1007/s41060-021-00302-z.

[25] J. Li, W. Jiang, J. Zhang, Y. Shao, and W. Zhu, "Fake User Detection Based on Multi-Model Joint Representation," *Information*, vol. 15, no. 5, pp. 266, 2024, doi: 10.3390/info15050266.

[26] M. Berrondo-Otermin and A. S. Cabezuelo, "Application of Artificial Intelligence Techniques to Detect Fake News: A Review," *Electronics*, vol. 12, no. 24, pp. 5041, 2023, doi: 10.3390/electronics12245041.

[27] A. A. Albraikan, M. Maray, F. A. Alotaibi, M. M. Alnfiai, A. Kumar, and A. E. Sayed, "Bio-Inspired Artificial Intelligence With Natural Language Processing Based on Deceptive Content Detection in Social Networking," *Biomimetics*, vol. 8, no. 6, pp. 449, 2023, doi: 10.3390/biomimetics8060449.

[28] M. Aljabri, R. Zagrouba, A. Shaahid, F. Alnasser, A. Saleh, and D. M. Alomari, "Machine Learning-Based Social Media Bot Detection: A Comprehensive Literature Review," *Soc. Netw. Anal. Min.*, vol. 13, no. 20, pp. 1–40, 2023, doi: 10.1007/s13278-022-01020-5.

[29] K. Shahzad, S. A. Khan, S. Ahmad, and A. Iqbal, "A Scoping Review of the Relationship of Big Data Analytics With Context-Based Fake News Detection on Digital Media in Data Age," *Sustainability*, vol. 14, no. 21, pp. 14365, 2022, doi: 10.3390/su142114365.

[30] K. Yang, E. Ferrara, and F. Menczer, "Botometer 101: Social Bot Practicum for Computational Social Scientists," *J. Comput. Soc. Sci.*, vol. 5, no. 2, pp. 1511–1528, 2022, doi: 10.1007/s42001-022-00177-5.

[31] A. Aldayel and W. Magdy, "Characterizing the Role of Bots' in Polarized Stance on Social Media," *Soc. Netw. Anal. Min.*, vol. 12, no. 30, pp. 1–24, 2022, doi: 10.1007/s13278-022-00858-z.

[32] Y. Zhang, W. Song, J. Shao, M. Abbas, J. Zhang, Y. H. Koura, and Y. Su, "Social Bots' Role in the COVID-19 Pandemic Discussion on Twitter," *Int. J. Environ. Res. Public Health*, vol. 20, no. 4, pp. 3284, 2023, doi: 10.3390/ijerph20043284.

[33] V. Suárez-Lledó and J. Álvarez-Gálvez, "Assessing the Role of Social Bots During the COVID-19 Pandemic: Infodemic, Disagreement, and Criticism," *J. Med. Internet Res.*, vol. 24, no. 8, pp. e36085, 2022, doi: 10.2196/36085.

[34] R. Kenny, B. Fischhoff, A. Davis, K. M. Carley, and C. Canfield, "Duped by Bots: Why Some Are Better Than Others at Detecting Fake Social Media Personas," *Hum. Factors J. Hum. Factors Ergon. Soc.*, vol. 66, no. 1, pp. 88–102, 2022, doi: 10.1177/00187208211072642.

[35] H. Y. Yan, K. Yang, J. Shanahan, and F. Menczer, "Exposure to Social Bots Amplifies Perceptual Biases and Regulation Propensity," *Sci. Rep.*, vol. 13, no. 1, pp. 1–10, 2023, doi: 10.1038/s41598-023-46630-x.

[36] S. Dalvi, S. Gholap, P. M. Jadhav, and P. P. Baviskar, "Social Media Fake Account Identification Using Ml," *Interantional J. Sci. Res. Eng. Manag.*, vol. 08, no. 11, pp. 1–7, 2024, doi: 10.55041/ijsrem38546.

[37] G. Shao, H. Zhang, S. Jin-yuan, K. Woeste, and L. Tang, "Strengthening Machine Learning Reproducibility for Image Classification," *Adv. Artif. Intell. Mach. Learn.*, vol. 02, no. 04, pp. 471–476, 2022, doi: 10.54364/aaiml.2022.1132.

[38] K. Cleal and D. M. Baird, "Dysgu: Efficient Structural Variant Calling Using Short or Long Reads," *Nucleic Acids Res.*, vol. 50, no. 9, pp. e53–e53, 2022, doi: 10.1093/nar/gkac039.

[39] C. Şahin, "Predicting Base Station Return on Investment in the Telecommunications Industry: Machine-learning Approaches," *Intell. Syst. Account. Financ. Manag.*, vol. 30, no. 1, pp. 29–40, 2023, doi: 10.1002/isaf.1530.

[40] M. Mendoza, E. Providel, M. Santos, and S. Valenzuela, "Detection and Impact Estimation of Social Bots in the Chilean Twitter Network," *Sci. Rep.*, vol. 14, no. 1, pp. 1–21, 2024, doi: 10.1038/s41598-024-57227-3.