Fake Profile Detection on Online Social Networks Using Machine Learning

Saptak Sil[1], Munmun Bhattacharya[2]

[1]PG Student, [2]Assistant Professor, [1,2]Department of Information Technology, Jadavpur University, Kolkata – 700106, India.

Abstract

Online Social Networks (OSNs) have become indispensable part for modern communication, commercial activity, and the dissemination of information. Platforms such as Twitter enable users to interact, share opinions, and engage with content at an unprecedented scale. However, this widespread adoption has also made OSNs attractive targets for malicious actors. Among the most pressing security concerns are profile cloning and fake engagement tactics that involve the creation and use of fake or automated accounts to manipulate popularity metrics, deceive genuine users, and distort public perception. These activities not only pose serious threats to individual privacy but also undermine trust in digital interactions, skew advertising strategies, and compromise the overall integrity of the platform. In this study, we focus on detecting such malicious behaviors on Twitter through the application of various machine learning algorithms. By conducting extensive experiments and performance evaluations, we compare traditional machine learning techniques with deep learning models. Our results demonstrate that deep learning approaches, particularly those leveraging neural network architectures, significantly outperform classical methods, achieving higher accuracy in identifying fake profiles and offering more robust solutions for OSN security.

Keywords: Online Social Networks, profile cloning, Twitter, machine learning, deep learning.

1. INTRODUCTION

Social media platforms like Instagram, Twitter and Facebook have become integral to daily life. These platforms serve a multitude of purposes, ranging from casual communication and community building to commercial transactions and brand promotion. However, the rise of fake accounts that spread misinformation and manipulate engagement metrics such as followers, likes, and comments poses serious risks. These deceptive practices can damage reputations and skew the perceived influence of users. In recent years, machine learning has emerged as a powerful tool in tackling this complex challenge by automating the detection of suspicious profiles based on patterns in user behavior, content characteristics, and network activity [3]. Online Social Networks (OSNs) face a multitude of challenges that compromise user experience, data integrity, and security [11]. Chief among these concerns are privacy violations, unauthorized data access, and the misuse of personal information, which can leave users vulnerable to identity theft, surveillance, and other malicious activities. Fake accounts and spamming reduce platform credibility and can be exploited for scams or misinformation. The rapid spread of rumors and the presence of abusive content contribute to misinformation and toxic online environments. Additionally, cyberbullying poses significant emotional risks, especially to younger users. These persistent issues highlight the urgent need for advanced detection mechanisms to build a better ecosystem. A comprehensive analysis of these threats and potential countermeasures is discussed in [18], emphasizing the role of technological interventions in strengthening OSN security. Moreover, in the current digital landscape, a growing number of users actively seek to boost their visibility or perceived influence on social media by resorting to unethical tactics. These include purchasing followers, likes, and comments, or employing automation tools and social bots to simulate engagement. While such practices may offer short-term gains, they erode the credibility of social metrics, mislead genuine users, and undermine the integrity of platform algorithms. Therefore, it is important to identify and understand fake accounts so that proper steps can be taken to stop such behavior. Social network detection techniques are broadly categorized into three primary approaches:

feature-based defense, graph-based defense, and hybrid techniques [10]. The feature-based defense approach primarily leverages behavioral analysis and profile attribute analysis. Behavioral analysis involves the examination of user activity patterns, such as posting frequency, login times, and interaction rates, as well as account characteristics like the age of the account and changes over time. Profile attribute analysis, on the other hand, focuses on static attributes including profile completeness, the presence of profile pictures, and similarities among various user profiles that may indicate automated or coordinated activity. In contrast, the graph-based defense approach focuses on the structural and topological properties of the social network. Additionally, trust propagation models are utilized, which apply algorithms like random walks or personalized PageRank to assess the trustworthiness of nodes within the network, allowing for the detection of anomalous or low-trust nodes that may signify fake or malicious accounts. Finally, hybrid techniques integrate feature-based and graph-based methodologies, combining the strengths of each approach to enhance overall detection accuracy and robustness. These hybrid models can leverage machine learning classifiers that process both attribute-level and structural features, providing a more comprehensive defense mechanism.
Our main contributions are as follows.

• We perform a comprehensive feature selection process to identify the most impactful attributes for detecting fake Twitter profiles from a set of 33 original features.

• We implement and evaluate multiple machine learning models, including an Artificial Neural Network (ANN), to detect fake accounts and assess their performance against baseline methods.

• We conduct a comparative analysis of our model's results with existing research that used the same dataset, highlighting the improvements in detection performance achieved through our approach.

The remainder of this paper is structured as follows: Section II provides the theoretical background, Section III outlines the proposed methodology, including data preprocessing, feature selection, and model architecture. The results of our experiments and their implications are presented in Section IV, followed by concluding remarks and potential directions for future work in Section V.

2. LITERATURE SURVEY

According to Fig. 1, several popular online social networks are illustrated. Among these platforms, Instagram has exhibited the most rapid growth, surpassing others in user engagement and activity.

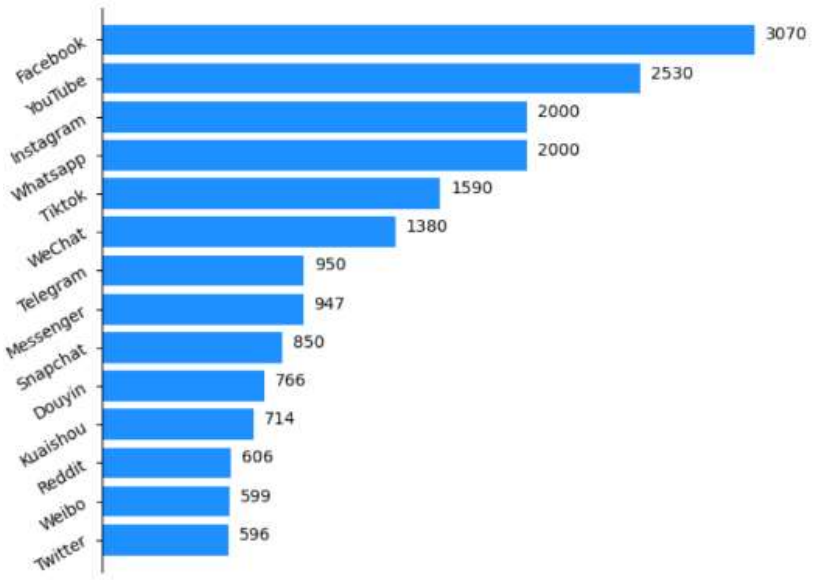

Fig. 1. Most Popular Social Networks Worldwide as of February 2025, [15]

Our study further reveals that the majority of academic and technical literature related to fake profile detection primarily focuses on Facebook, Instagram, and Twitter. These platforms dominate the research landscape due to their expansive user bases, openness of certain APIs, and frequent use by malicious actors for spreading misinformation, conducting phishing attacks, and creating fraudulent or sybil accounts. A careful review and analysis of relevant studies shows that researchers predominantly employed machine learning (ML) algorithms to address the challenge of detecting fake users, profile cloning and sybil attacks. In particular, supervised learning models such as Decision Trees, Random Forests, Support Vector Machines (SVM), and Artificial Neural Networks (ANNs) are commonly used for classification of genuine versus fake profiles.

Table 1 Summary of Related Work

Author	Dataset Used	Public Dataset	Dataset Type
Kadam et al. [7]	Twitter	Yes	Twitter tweets Dataset
Erşahin et al. [5]	Twitter	No	Twitter User Metadata and Recent Tweet Behavior Dataset
Akyon et al. [1]	Instagram	Yes	Instagram User Dataset
Tunç, Ümmü et al. [17]	Instagram	Yes	Instagram User Profile and Engagement Dataset
Purba et al. [9]	Instagram	Yes (Kaggle)	Instagram User Behavioral and Content Feature Dataset
Sonowal et al. [14]	Instagram	Yes (Kaggle)	Instagram User Dataset
Ekosputra et al. [4]	Instagram	Yes (Kaggle)	Instagram User Dataset
Gupta et al. [6]	Facebook	No	Facebook User Interaction and Engagement Dataset
Shreya et al. [12]	Facebook	No	Facebook User Interaction
Narayanan et al. [8]	Twitter	No	Twitter User Metadata Dataset
Suriakala et al. [16]	GitHub	No	GitHub User Account Dataset
Suriakala et al. [2]	Twitter	No	Twitter Behavioral and Interaction Dataset

Table 1 summarizes a curated selection of research papers that have made significant contributions in this domain. These studies span multiple platforms, including Twitter, Instagram, Facebook, and GitHub, reflecting the diverse landscape of OSNs and the varying challenges posed by each. The research community has explored a wide array of methodologies for identifying and analyzing inauthentic behavior, ranging from traditional machine learning classifiers to more recent deep learning and graph-based approaches. These works span various social media platforms and employ diverse datasets some publicly available and others private to analyze different types of fake behavior. Although, researchers have made substantial progress, but challenges such as dataset availability, evolving evasion tactics, and platform restrictions still present open problems in the field. This paper focuses on detecting fake profile accounts to increase the trustworthiness of OSNs.

3. PROPOSED WORK

To develop an effective machine learning model for the detection of fake profiles in the early stages, we have designed an experimental framework. The structure of this proposed model is illustrated in Fig. 2, and its components and functionality are explained in detail in the following section.

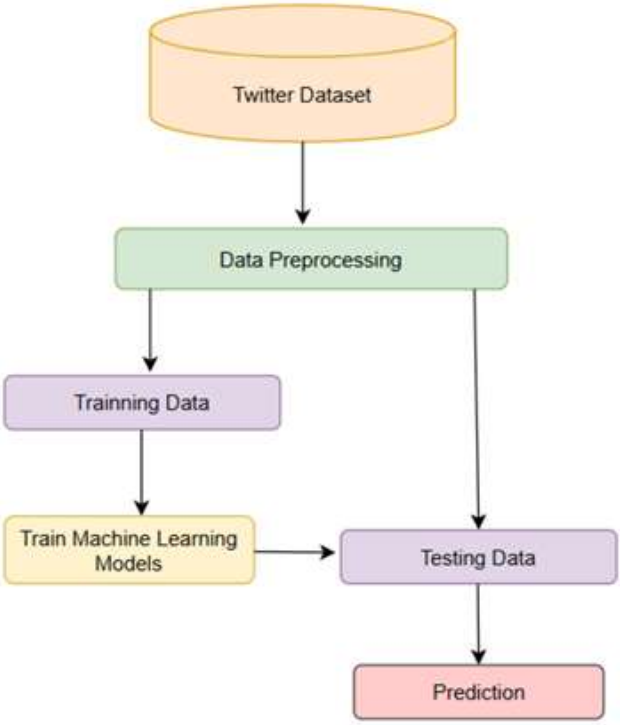

Fig. 2. Proposed Model

3.1 Dataset

The Dataset is available at [7]. The dataset consists of 33 attributes. The fake and genuine users are present in two different CSV files. The distribution of both user types is illustrated in Fig. 3. After merging the two files and labeling the instances accordingly, the combined dataset consists of 2,820 entries.

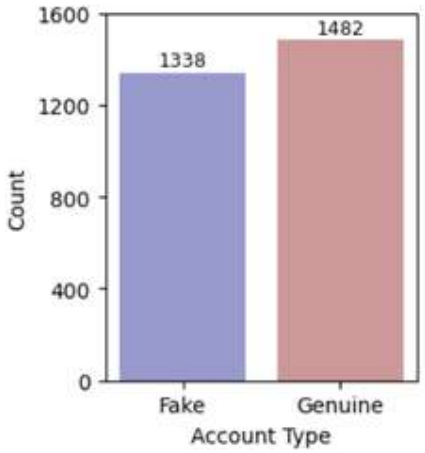

Fig. 3. Account type distribution in the dataset

3.2 Dataset Analysis

The original dataset comprised 33 features capturing various aspects of Twitter user profiles. However, not all features contributed equally to the task of distinguishing between genuine and fake users. An initial examination revealed that there were no missing (null) values in the dataset. The features were primarily of three data types: integer, float, and object. Among the object-type features, notable ones include created at, description, and time zone. While these are stored as objects, their semantics differ:

- created at - represents the date the user account was created and can be converted to a datetime format for temporal analysis.
- description - contains the user-defined textual description of the profile, which may be useful for natural language processing or keyword-based features.
- time zone - indicates the geographic time zone associated with the account, which could offer regional insights.

3.3 Dataset Preprocessing

A feature selection process was carried out to retain only the most relevant attributes for model training. Features like id, name, screen name do not contribute in differentiating between fake and real users. Key attributes like statuses count, followers count, friends count, favourites count, and listed count serve as core indicators of user behavior. In contrast, features like URL and lang are considered noninformative for this task. Several features were binarized to ensure compatibility with the learning model. Boolean fields like geo_enabled, default profile and profile use background image were directly converted to binary. Fields such as description and time zone were encoded as 1 if a value was present and 0 if left blank or missing. Although protected and verified samples could be informative, they were excluded because of the presence of only NaN values across all records. All numerical features were normalized using Min-Max normalization to bring them to a consistent scale as the data was unevenly distributed. Since there were not many outliers, normalization was preferred over standardization. Duplicate records were removed, resulting in a final dataset of 1,481 genuine users and 1,329 fake users. After normalization we noticed that the updated column which was having count of days, the profile was last updated had all the count to zero. As a result, this feature was removed due to its lack of variance, while other non-essential attributes were omitted to maintain focus. Ultimately, we retained 12 essential features for model training. A summary of the final features is shown in Table 2.

Table 2 Final Feature Set Used for Classification

Features	Transformation
followers_count	Normalized (Min-Max)
friends_count	Normalized (Min-Max)
statuses_count	Normalized (Min-Max)
favourites_count	Normalized (Min-Max)
listed_count	Normalized (Min-Max)
created_at	Converted to days and then normalized
geo_enabled	if present 1 else 0
default_profile	if present 1 else 0
profile_use_background_image	if present 1 else 0
description	if non-empty 1 else 0
profile_background_tile	if present 1 else 0
time_zone	if present 1 else 0

3.4 Pattern Learning and Classification

The preprocessed data was divided into two sets: 70% for training and 30% for testing. Several machine learning algorithms were applied, along with an Artificial Neural Network (ANN), which demonstrated particularly strong performance. The ANN was built using a sequential model structure that included four hidden layers with 128, 64, 32 and 16 neurons respectively, each using the ReLU activation function to introduce non-linearity. The final output layer consisted of a single neuron with a sigmoid activation function, which is ideal for binary classification tasks. The model was compiled using the Adam optimizer and binary cross-entropy as the loss function, with accuracy as the evaluation metric. We trained the model for 100 epochs with a batch size of 32, using 10% of the training data for validation to monitor the model's performance. After training, the model predicted probabilities for the test data, which were then converted into binary class labels: if the predicted probability was greater than 0.5, the user was classified as real (label 0); otherwise, the user was classified as fake (label 1). The details of the neural network are given in Table 3.

Table 3 NEURAL NETWORK DETAILS

Parameter	Value
No of Layers	4 Hidden layer and 1 output layer
No of Hidden Units (per layer)	128, 64, 32, 16
Optimization	ADAM
Non-linearity	ReLU (used in all hidden layers)
Loss Function	binary Cross-Entropy
Learning Rate	0.001 (Default)
Minibatch Size	32
Epochs	100
Train-Test Split	70% - 30%

The steps of detection of the fake profiles are summarized as follows:
1) Collect the Twitter dataset from [7].
2) Split the dataset into train and test data.
3) Train the model.
4) Predict the test result of fake and genuine profiles.
5) Compute accuracy.
6) Create the Confusion matrix for classification and misclassification of Twitter profiles.
7) Compare the performance of classification models using ROC.

4. RESULTS

The primary objective of this paper is to assess the effectiveness of various classification algorithms in detecting fake users on Twitter. To evaluate their performance, standard machine learning metrics are employed. These include True Positives (TP), representing correctly identified real profiles; True Negatives (TN), denoting accurately classified fake profiles; False Positives (FP), where fake profiles are incorrectly identified as real; and False Negatives (FN), where real profiles are mistakenly classified as fake. All experiments were conducted using the Python programming language within the Google Colab environment, chosen for its flexibility and ease of access. Google Colab offers a cloud-based platform that facilitates code execution without the need for local configuration. Moreover, our models run efficiently even on Colab as standard CPU runtime.

4.1 Accuracy

Accuracy represents how effectively a classification algorithm correctly predicts outcomes. It is calculated as the ratio of the number of correct predictions to the total number of predictions made. This fundamental metric provides a straightforward and intuitive means to evaluate the overall performance of a model in distinguishing between different classes. The mathematical formulation for accuracy is given as:

$$\text{Accuracy} = TP + TN / TP + TN + FP + FN \quad (1)$$

The accuracy of all the evaluated algorithms is presented in Fig 4. From the comparison, it is evident that the Artificial Neural Network achieved the highest accuracy among the classifiers considered. This indicates the superior capability of ANN in correctly identifying both genuine and fake profiles compared to other traditional classifiers such as Support Vector Machine, Logistic Regression, and Naive Bayes. Furthermore, the figure illustrates not just the accuracy but also additional performance metrics such as Precision, Recall, and F1-score for each classifier. These metrics provide a more comprehensive evaluation by highlighting how well the models balance between false positives and false negatives. ANN consistently outperforms across all metrics, suggesting its robustness and reliability in detecting fake accounts, which is crucial for security-sensitive applications.

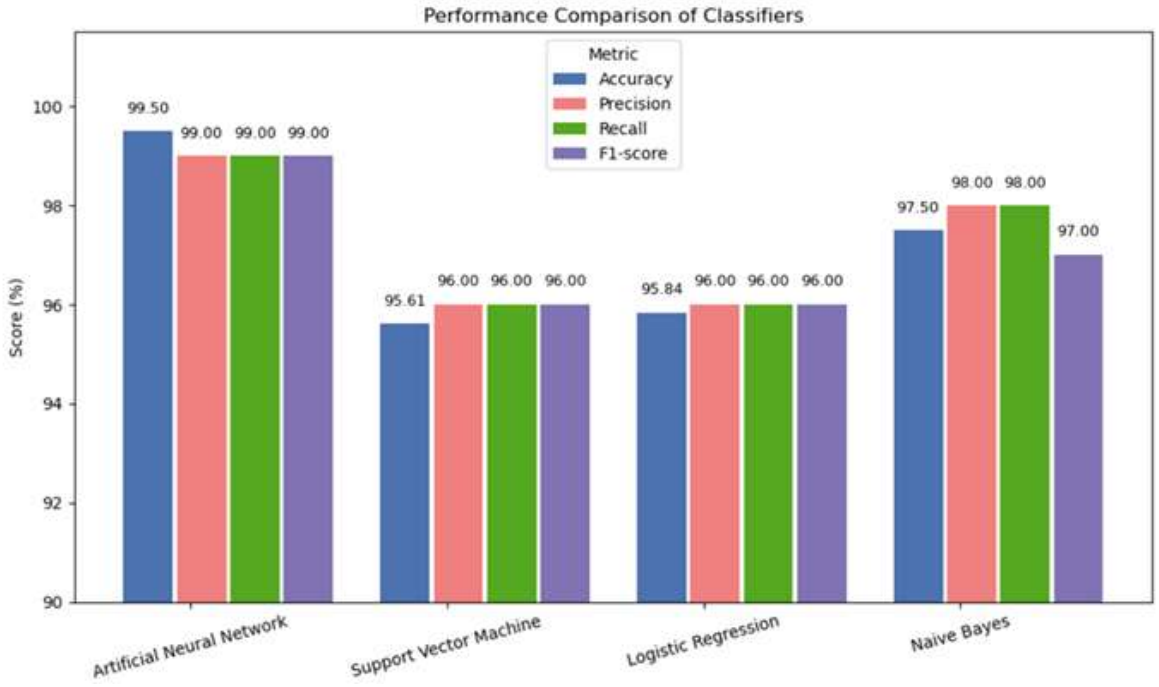

Fig. 4. Performance Comparison of Classifiers

The confusion matrix of the Artificial Neural Network (ANN), shown in Fig. 5, provides a detailed representation of the model's effectiveness in distinguishing between Genuine (0) and Fake (1) profiles. Out of 450 genuine instances, the model accurately classified 446, misclassifying only 4 as fake, thereby exhibiting a very low false positive rate. More impressively, the model achieved perfect classification of all 393 fake profiles, resulting in zero false negatives. This leads to a recall of 100% for the fake class, which implies that the model was able to correctly identify every single fake account present in the dataset. This results in a 100% recall for the fake class, indicating that every fake account was accurately identified by the model. In practical terms, this means that if the algorithm flags an account as fake, it is almost certainly fake, which is highly desirable in applications requiring robust detection of malicious profiles.

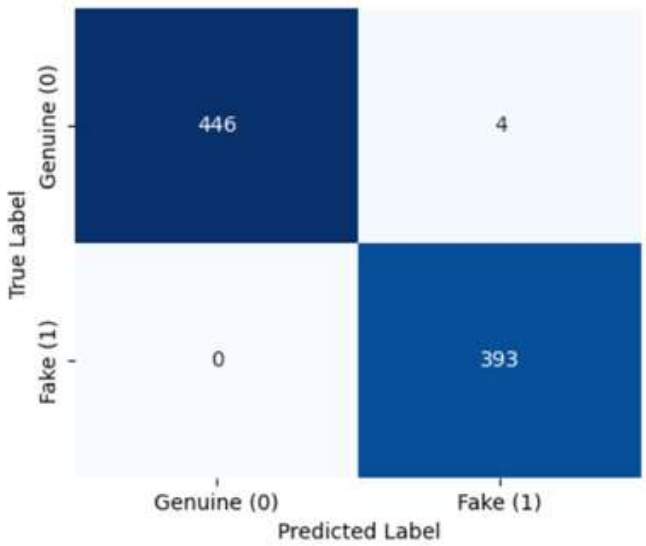

Fig. 5. Confusion Matrix of ANN

4.2 Time Consumption

Time consumption is a critical parameter to consider during model selection, especially in real-time or resource-constrained environments. It refers to the total duration taken by a model to complete its training or execution, typically calculated as the difference between the start and end time of the process. Although accuracy and performance metrics are often prioritized in evaluating models, computational efficiency plays an equally important role particularly when deploying models in production environments. Models with high accuracy but excessive training times may not be practical for real-world applications. Therefore, balancing predictive performance with computational efficiency is essential for selecting optimal models, especially when working with large datasets or integrating into time-sensitive systems.

$$\text{Time Consumed} = \text{End Time} - \text{Start Time} \quad (2)$$

Table 4 presents the time consumption of different algorithms. Naive Bayes exhibits the lowest time consumption, outperforming the other algorithms in terms of computational efficiency.

Table 4. Time consumption (in milliseconds)

Classifier	Time Consumed (mS)
Artificial Neural Network **(**ANN)	30682.73
Logistic Regression (LR)	84.28
Support Vector Machine (SVM)	48.51
Naive Bayes **(**NB)	9.80

4.3 ROC Curve

The ROC curve provides a comprehensive summary of model performance across all classification thresholds by capturing the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR). As shown in Fig. 6, all four models ANN, SVM, Logistic Regression, and Naive Bayes exhibit excellent classification performance. Their ROC curves closely follow the top left edge of the plot, indicating high TPRs and low FPRs throughout. ANN achieves a perfect AUC of 1.00, suggesting it separates the classes flawlessly on the test set. This is consistent with its confusion matrix, which reflects a very high recall. However, ANN is also the most computationally intensive model. SVM and Naive Bayes offer near-equal

performance with less complexity. Logistic Regression is slightly less effective but more interpretable and faster to train. Since all AUCs are above 0.98 and the curves overlap closely, the dataset might be easy to classify, or the features used are highly informative.

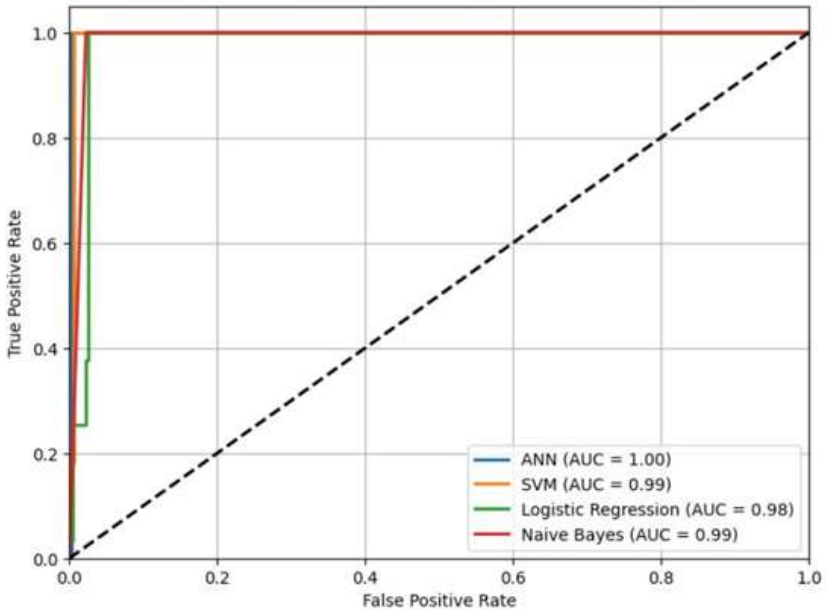

Fig. 6. ROC Curve Comparison Across Models

4.3 Comparison Analysis

We conducted a comparative analysis between our best-performing model and the top-performing approach proposed by Kumar et al. The results, presented in Table 5, demonstrate that our Artificial Neural Network (ANN) model outperforms their method, achieving a higher accuracy of 99.5% compared to 97.4%.

Table 5. Comparison analysis between the proposed method and existing work

Author	Classifier	Accuracy (%)	Time Consumed (mS)
Kumar et al.	ANN	97.4	365
Our Work	ANN	99.5	30682.73
	NB	97.5	9.8

Additionally, it is worth noting that the Naive Bayes classifier also shows competitive performance, achieving an accuracy of 97.5%. Furthermore, Naive Bayes exhibits the lowest time consumption among all evaluated models, making it a strong candidate in terms of computational efficiency and accuracy compared to existing methods.

5. CONCLUSION

In this study, our objective was to distinguish between fake and real Twitter accounts using various machine learning algorithms. We employed several models, including Artificial Neural Networks (ANN), Logistic Regression (LR), Support Vector Machines (SVM), and Naive Bayes (NB). Among these, ANN achieved the highest accuracy. However, it also incurred the highest computational cost, with significantly longer training times compared to the others. On the other hand, the Naive Bayes model provided a well-balanced trade-off between accuracy and computational efficiency. Additionally, we performed 4-fold cross-validation, which yielded an impressive accuracy of 98.4%. While the current study demonstrates promising results in detecting fake Twitter accounts using machine learning techniques, there are several

avenues for future improvement and exploration such as incorporating additional features such as tweet content analysis, temporal patterns and user interactions.

6. CODE AVAILABILITY

The source code is made publicly available in [13].

REFERENCES

[1] Akyon, F. C., & Kalfaoglu, M. E. (2019, October). Instagram fake and automated account detection. In 2019 Innovations in intelligent systems and applications conference (ASYU) (pp. 1-7). IEEE.

[2] Alsaleh, M., Alarifi, A., Al-Salman, A. M., Alfayez, M., & Almuhaysin, A. (2014, December). Tsd: Detecting sybil accounts in twitter. In *2014 13th International Conference on Machine Learning and Applications* (pp. 463-469). IEEE.

[3] Bhattacharya, M., Roy, S., Chattopadhyay, S., Das, A. K., & Shetty, S. (2023). A comprehensive survey on online social networks security and privacy issues: Threats, machine learning-based solutions, and open challenges. *Security and Privacy*, *6*(1), e275.

[4] Ekosputra, M. J., Susanto, A., Haryanto, F., & Suhartono, D. (2021, December). Supervised machine learning algorithms to detect instagram fake accounts. In *2021 4th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)* (pp. 396-400). IEEE..

[5] Erşahin, B., Aktaş, Ö., Kılınç, D., & Akyol, C. (2017, October). Twitter fake account detection. In *2017 international conference on computer science and engineering (UBMK)* (pp. 388-392). IEEE.

[6] Gupta, A., & Kaushal, R. (2017, January). Towards detecting fake user accounts in facebook. In *2017 ISEA Asia Security and Privacy (ISEASP)* (pp. 1-6). IEEE.

[7] Kadam, N., & Sharma, S. K. (2022). Social media fake profile detection using data mining technique. *Journal of Advances in Information Technology*, *13*(5), 518-523.

[8] Narayanan, A., Garg, A., Arora, I., Sureka, T., Sridhar, M., & Prasad, H. B. (2018, December). IronSense: towards the identification of fake user-profiles on twitter using machine learning. In *2018 Fourteenth International Conference on Information Processing (ICINPRO)* (pp. 1-7). IEEE.

[9] Purba, K. R., Asirvatham, D., & Murugesan, R. K. (2020). Classification of instagram fake users using supervised machine learning algorithms. *International Journal of Electrical and Computer Engineering*, *10*(3), 2763.

[10] Ramalingam, D., & Chinnaiah, V. (2018). Fake profile detection techniques in large-scale online social networks: A comprehensive review. *Computers & Electrical Engineering*, *65*, 165-177.

[11] Roy, P. K., & Chahar, S. (2021). Fake profile detection on social networking websites: a comprehensive review. *IEEE Transactions on Artificial Intelligence*, *1*(3), 271-285.

[12] Shreya, K., Kothapelly, A., & Shanmugasundaram, H. (2022, December). Identification of Fake accounts in social media using machine learning. In *2022 Fourth International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT)* (pp. 1-4). IEEE.

[13] Sil, S. (2025). *Profile cloning: Fake profile detection using machine learning* [Source code]. GitHub. https://github.com/silsaptak/Profile_Cloning

[14] Sonowal, G., Balaji, V., & Kumar, N. (2024). A model to detect fake profile on instagram using rule-based approach. *CSI Transactions on ICT*, *12*(4), 95-105.

[15] Statista. (2025, February). *Most popular social networks worldwide as of February 2025, by number of monthly active users*. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

[16] Suriakala, M., & Revathi, S. (2018, December). Privacy protected system for vulnerable users and cloning profile detection using data mining approaches. In *2018 Tenth International Conference on Advanced Computing (ICoAC)* (pp. 124-132). IEEE.

[17] Tunç, Ü., Atalar, E., Gargı, M. S., & Aydın, Z. E. (2022). Classification of fake, bot, and real accounts on instagram using machine learning. *Politeknik Dergisi*, *27*(2), 479-488.

[18] Zhang, C., Sun, J., Zhu, X., & Fang, Y. (2010). Privacy and security for online social networks: challenges and opportunities. *IEEE network*, *24*(4), 13-18.