

# Multi platforms fake accounts detection based on federated learning

Marina Azer<sup>1</sup>, Hala H. Zayed<sup>1,2</sup>, Mahmoud E. A. Gadallah<sup>3</sup>, Mohamed Taha<sup>1</sup>

<sup>1</sup>Faculty of Computers and Artificial Intelligence, Benha University, Benha, Egypt

<sup>2</sup>Faculty of Engineering, Egypt University of Informatics, Cairo, Egypt

<sup>3</sup>Faculty of Computer Science, Modern Academy, Cairo, Egypt

## Article Info

### Article history:

Received Feb 12, 2024

Revised Apr 20, 2024

Accepted Jun 14, 2024

### Keywords:

Data privacy

Fake account detection

Federated learning

Machine learning

Social media

## ABSTRACT

Identifying and mitigating fake profiles is an urgent issue during the age of widespread integration with social media platforms. this study addresses the challenge of fake profile detection on major social platforms-Facebook, Instagram, and X (Twitter). Employing a two-sided approach, it compares stacking model of machine learning algorithms with the federated learning. The research extends to four datasets, two Instagram datasets, one X dataset, and one Facebook dataset, reporting impressive accuracy metrics. Federated learning stands out for its effectiveness in fake profile detection, prioritizing user data privacy. Results reveal Instagram fake/real dataset achieves 96% accuracy while Instagram human/bot dataset reaches 95% accuracy with federated learning. using the stacking model X's fake/real dataset achieves 99.4% accuracy, and Facebook fake/real dataset reaches 99.8% accuracy using the same model. The study underscores the pivotal role of data privacy, positioning federated learning as an ethical choice. It compares the time efficiency of stacking and federated learning, with the former providing good performance in less time and the latter emphasizing data privacy but consuming more time. Results are benchmarked against related works, showcasing superior performance. The study contributes significantly to fake profile detection, offering adaptable solutions and insights.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



## Corresponding Author:

Marina Azer

Faculty of computers and Artificial intelligence, Benha University

Benha, Egypt

Email: marina.essam991@gmail.com

## 1. INTRODUCTION

Social media's rapid expansion presents both opportunities and challenges for marketing campaigns and celebrity promotion. Fake profiles, often created to misrepresent individuals or entities, can damage reputations and distort engagement metrics. Moreover, they contribute to confusion and facilitate cyberbullying. Privacy concerns vary among users in the online realm. With the proliferation of big data platforms like social media, identity fraud has become a pressing issue. Social media platforms have become prime targets for spammers and con artists, posing various threats. Instagram stands out as a dominant social networking platform, but alongside its popularity comes increased exploitation of users. X (Twitter), known for its simplicity, hosts a mix of real users and bots, with the latter often masquerading as humans. Facebook, as a leading social networking platform, faces similar challenges with the proliferation of fake profiles and the exploitation of users. Detecting social media bots is crucial for maintaining online discourse integrity, prompting extensive research into relevant datasets. Social bots, or sybil accounts, are automated algorithms

that engage with users on social media. While some are harmless or even entertaining, others are used for deceptive purposes, such as creating fake grassroots support for political agendas. Addressing the issue of fake human accounts on social media platforms requires a multifaceted approach, including machine learning algorithms for detection and stricter verification procedures. Resolving this problem is essential for ensuring a safe and trustworthy online environment for all users [1]–[3].

The objective of this study is to develop and evaluate effective strategies for identifying and mitigating fake profiles within these diverse and dynamic online environments. In the pursuit of this goal, we employ a two-fold approaches. Firstly, we harness the power of machine learning methodologies, leveraging various algorithms to construct a robust ensemble model. This model is designed to enhance the ability to differentiate between genuine and fraudulent profiles [4]. Secondly, we explore the innovative paradigm of federated learning. Federated learning presents an intriguing opportunity to address fake profile detection while respecting the paramount importance of data privacy to consider the sensitive nature of the data. federated learning offers a compelling solution to the ethical and technical challenges of fake profile detection in a privacy-aware world by enabling model training without centralized access to sensitive user data. Offers a comparative analysis between the ensemble machine learning model and the federated learning approach across the four datasets spanning the three platforms. This comparison scrutinizes their performance in terms of accuracy metrics and time consumption [5].

The challenges inherent in this work stem from various factors. Firstly, the reliability and representativeness of the datasets utilized greatly influence the validity of the study's findings. Secondly, the substantial computational resources demanded for training and assessing intricate machine learning models, particularly in the context of federated learning, present practical limitations. These constraints could hinder the scalability and real-world applicability of the proposed approach, moreover, external factors, including alterations in platform regulations, shifts in user behavior, and advancements in technology, may impact the effectiveness of the proposed methodology over time. Thus, ongoing monitoring and adaptation are imperative to navigate these external influences and ensure the continued relevance and efficacy of the approach.

This study contributes to the existing body of research by investigating the effects of fake account detection using two different approaches on three different platforms. Most earlier studies have not explicitly addressed the influence of fake account detection across multiple platforms always focus on specific platform and have not focused on data privacy approaches in this topic. Therefore, by examining the efficacy of detection methods across various platforms and emphasizing the importance of data privacy, this research expands our understanding of fake profile detection strategies and their implications in diverse online environments. Sarhan and Mattar [6] compare machine learning algorithms performance to categorize user accounts on the Instagram platform, Rostami [7] focused on feature selection to enhance classification accuracy. Utilizing a multi-objective hybrid feature selection approach, the researchers identify a candidate feature set with high relevance to the target class and minimal redundancy using the minimum redundancy maximum relevance (mRMR) algorithm. The final feature set is chosen for detection operations, yielding superior performance compared to existing methods when tested on Twitter datasets. Hakimi *et al.* [8] established a comprehensive set of five critical characteristics pivotal in discerning fake from genuine users on Facebook. Subsequently, we employ these attributes to identify key classifiers in machine learning, with particular focus on K-nearest neighbor (KNN), support vector machine (SVM), and neural network (NN). Results indicate that KNN emerges as the top performer, achieving an 82% classification precision rate among the classifiers evaluated.

In conclusion, this research is poised to significantly contribute to the ongoing efforts to fight fake profiles on social media. By using machine learning techniques and the innovative privacy-preserving capabilities of federated learning, we aim to fortify the security and trustworthiness of online interactions while supporting the principles of data privacy in today's digital landscape. The structure of this paper is as follows: section 2 outlines the research methodology. Section 3 presents the results and discussion. Finally, section 4 concludes the paper and providing suggestions for future research areas.

## 2. METHOD

The proposed model for fake profile detection, as shown in Figure 1, takes a systematic, multi-phased approach to achieve robust results. It begins with comprehensive data gathering from three major social media platforms: Facebook, Instagram, and X (Twitter). This process entails acquiring user profiles, their associated activities, and any relevant metadata. Subsequently, meticulous data preprocessing techniques are employed to clean and ensure consistency within the data. Feature engineering serves as a crucial step where domain-specific features are carefully crafted to capture the subtle behavioral patterns that differentiate fake profiles from genuine ones. To facilitate model evaluation on unseen data, the data is then strategically split into training and testing sets. Following this, the model diverges into two paths, offering a

comprehensive approach: several machine learning algorithms leverage the engineered features to construct predictive models by training the model on the training data and make a predictions on the test set ,then in the meta\_model's training phase the model is functioning as a stacking model which is designed to analyze and combine the predictions generated by the five most accurate machine learning models The meta-model's output is then considered the final prediction of the system in the first path, while the second path explores federated learning as a privacy-preserving alternative, enabling model training without having to share the data itself and without centralizing user data from the three platforms.

The proposed model for fake profile detection entails navigating through various complexities inherent in the process. Firstly, comprehensive data gathering from major social media platforms like Facebook, Instagram, and X (Twitter) presents a significant challenge due to differing data formats. Following data collection preprocessing techniques are essential to ensure data quality and consistency, involving tasks such as missing values handling and normalization. Moreover, crafting domain-specific features for effective differentiation between fake and genuine profiles adds another layer of complexity to the model. Selecting relevant features and transforming raw data into meaningful inputs demands a deep understanding of social media behavior and deception tactics. Subsequently, evaluating the performance of machine learning models on unseen data becomes crucial, necessitating careful selection of evaluation metrics and strategies to avoid overfitting. Finally, the implementation of the federating learning approach across multiple platforms involves technical challenges such as compatibility issues, scalability concerns, and privacy considerations. Despite these difficulties, the model aims to achieve robust and effective fake profile detection capabilities.

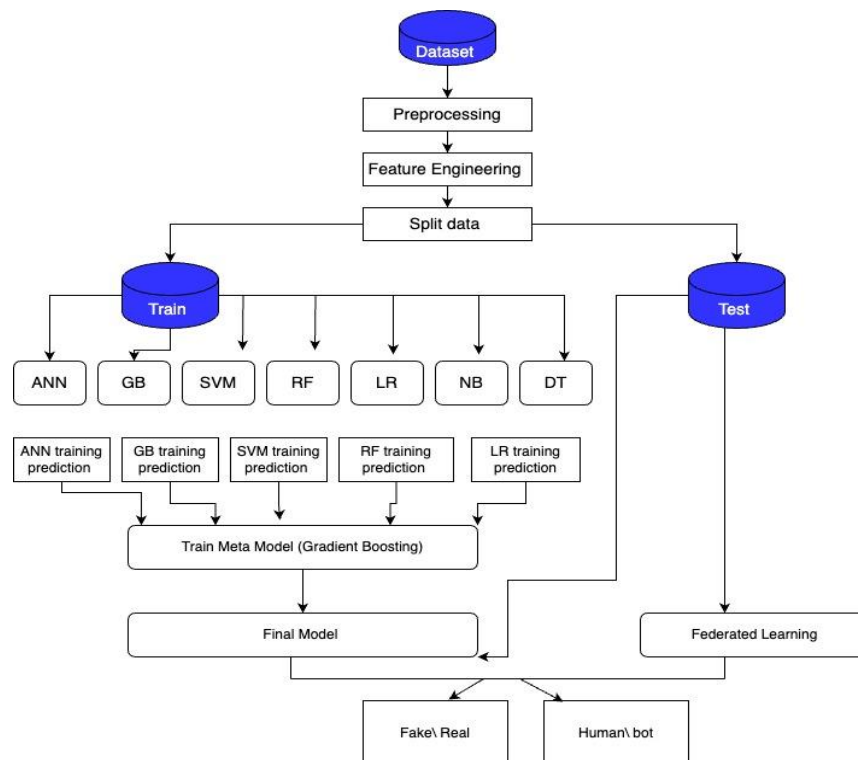


Figure 1. The proposed model

## 2.1. Data collection

For this comprehensive research on fake profile detection, we meticulously curated and utilized four distinct datasets, each sourced from the prominent social media platforms of Instagram, Twitter, and Facebook. These datasets collectively represent a diverse and comprehensive cross-section of user behaviors and profiles from the ever-evolving social media landscape. The Instagram dataset encapsulates the dynamic visual interactions of users, while the Twitter dataset brings forth the realm of concise and rapid-fire communication. The Facebook dataset provides insights into the multifaceted activities of users within this expansive social network. Collectively, these datasets enable a comprehensive analysis of fake profile detection across a spectrum of platforms, facilitating a robust evaluation of the proposed detection models.

### 2.1.1. The Instagram datasets

In this study, two diverse datasets were employed to comprehensively analyze profiles on Instagram. Firstly, we utilized the “Instagram fake accounts dataset” sourced from Kaggle, it is named Dataset 1. This dataset encompasses a total of 786 Instagram profiles, meticulously categorized into 693 fake accounts and 94 authentic profiles. These profiles are accompanied by a rich set of 12 features. Furthermore, our analysis extends to a second dataset, denoted as Dataset 2, which specifically targets the detection of bot-driven activity on Instagram. This dataset comprises a balanced set of 700 real accounts alongside 700 automated accounts, totaling 1,400 profiles. Each profile is characterized by a comprehensive set of 16 features.

### 2.1.2. The X (Twitter) dataset

In addition to our exploration of Instagram profiles, the study encompasses an investigation into Twitter account behavior using the ‘MIB’ dataset (Dataset 3) This dataset amalgamates three distinct datasets compiled from Twitter accounts, comprising a total of 5,301 profiles. Among these, 3,351 profiles are flagged as fake accounts, while 1,950 profiles are authenticated as genuine. Each profile in this dataset is associated with a detailed set of 34 features, providing insights into various aspects of account activity, including tweet frequency, retweet behavior, follower demographics, and linguistic patterns.

### 2.1.3. The Facebook dataset

Obtaining a dataset, Dataset 4, for studying Facebook user profiles is challenging due to privacy and access restrictions. However, this research acquired a dataset consisting of 2,820 profiles. Among these, 1,482 were confirmed as real users, while 1,338 were suspected to be fake accounts. The dataset encompasses 34 distinct features, making it a valuable resource for investigating the complexities of fake account detection on Facebook. This highlights the importance of ethical data collection practices and the significance of such datasets in research endeavors in this domain.

## 2.2. Preprocessing

The preprocessing phase in our research is a crucial step that adapts to each dataset’s unique characteristics and idiosyncrasies sourced from Instagram, Twitter, and Facebook. This phase ensures the data is standardized and cleaned, setting the stage for subsequent analysis. Throughout each dataset’s preprocessing journey, we meticulously addressed missing data, outliers, and inconsistencies, thus preparing the data for the subsequent feature engineering and model training steps.

In the context of our Instagram fake and real dataset, it became evident that we were dealing with imbalanced data, where the number of fake and real profiles demonstrated a significant variation. To address this issue and ensure the robustness of our analysis, we employed an oversampling technique known as synthetic minority over-sampling technique for nominal and continuous features (SMOTE-NC) [9]. This technique was chosen due to its demonstrated efficacy in situations involving imbalanced data comprising numerical and categorical features [10].

In the Twitter dataset, check null values using the null function, outliers, and class balance and convert categorical features to numeric values. In the Facebook dataset, there are 34 features; some of them, like verified and protected, are empty without values, so they are dropped, geo-enabled, default profile image, profile\_background\_title, utc-offset, and time zone have just 1% values of all users so here the imputation is not a good solution, so they are also dropped. There were null values in other features imputed using mean with numeric values.

### 2.2.1. Feature engineering

The feature engineering phase is a crucial step in the data preprocessing pipeline, where raw data undergoes transformation and manipulation to create new, informative features that enhance the performance of machine learning models. This phase involves extracting relevant information from the available dataset, selecting and combining features, and generating new representations that capture meaningful patterns and relationships within the data. Feature engineering aims to improve the predictive power of models by refining the input data [11].

#### – Instagram fake-real dataset

The feature engineering phase encompasses the utilization of 12 distinct features to analyze engagement dynamics and profile authenticity. These features include parameters such as “edge followed by,” “edge follow,” “user name length,” “user name containing a number,” “full name containing a number,” “full name length,” “is private,” “is joined recently,” “has a channel,” “is a business account,” “has guides,” and “has an external URL”. Through calculating the ratio of followers to accounts followed, potential imbalances in follower-following relationships can be detected, aiding in the identification of suspicious accounts. Additionally, techniques like Jaccard similarity [12] are employed to assess similarities between

usernames and full names, with notable differences potentially signaling accounts warranting further investigation. A composite feature is generated, amalgamating “is private,” “description length,” and “has external URL,” offering insights into the completeness of bio information, where multiple links in the description could raise suspicion. Further analysis involves calculating interaction ratios, numeric ratios between names and usernames, examining non-alphanumeric characters, name and username length ratios, and guides-to-posts ratios. Elevated ratios may signify suspicious behavior. Moreover, features indicating username similarity to verified accounts and the presence of associated external channels or platforms are created. Fake accounts are less likely to be connected to external channels, often prioritizing spam promotion, malicious activities, or user impersonation. Through comprehensive feature engineering, these parameters collectively contribute to the robust identification and characterization of fraudulent profiles.

– Instagram bot/human dataset

The dataset encompasses 16 distinct features aimed at capturing various aspects of user and media behavior on the platform. These features include metrics such as “media-like numbers,” “media comment numbers,” “media comments disabled,” “media hashtag numbers,” “media upload times,” “media has location info,” “user follower count,” “user following count,” “user has high reels,” “user has external URL,” “user tags count,” “user biography length,” “user name length,” “user name digit count,” “automated behavior,” and “user media count”. These features serve as the foundation for computing essential statistics such as average likes per media, average comments per media, and the ratio of likes to comments. Additionally, the dataset facilitates the analysis of posting time variability across all media, the frequency of media uploads over a specified period, and the diversity of hashtags and locations. Furthermore, key ratios such as follower-to-like ratio and follower-to-comment ratio can be calculated, shedding light on the engagement dynamics and audience interactions associated with each user. By leveraging these comprehensive features, we can gain valuable insights into user behavior patterns and media engagement trends, ultimately enabling more informed decision-making and analysis on the platform.

– Twitter fake/real dataset

Twitter data consists of 34 features, which are (profile link color, profile background color, profile sidebar fill color, profile background title, profile banner URL, profile text color, universal time coordinated (UTC) offset, default profile image, default profile, geo-enabled, listed count, favorites count, friends count, followers count, statuses count, profile background image URL https, profile sidebar border color, screenname, protected, verified, description, updated, dataset, created at, URL, lang, time zone, location, profile image URL, name, id, profile image URL https, profile background image URL and profile use background).because of the high number of features, essential features should be selected. By utilizing various base models such as logistic regression (LR), random forest (RF), SVM, gradient boosting (GB), and artificial neural network (ANN) on the dataset to extract feature importance scores.

Each model will evaluate and assign scores to individual features based on its perception of their significance in identifying fake profiles. This process allows for a comprehensive understanding of feature relevance across different modeling approaches. The typical range for importance scores lies between 0 and 1, where 0 signifies features of low or negligible importance, while 1 denotes features of high significance in detecting fake profiles. This standardized scale facilitates the interpretation of feature importance across different models and datasets based on the aggregated importance scores obtained from various models, the top twenty selected features [13] as they are likely to offer the most discriminative power in distinguishing between fake and genuine profiles are identified as follows: high importance (score: 0.6-1) profile link color, profile background color, profile sidebar color, profile banner URL, profile text color, default profile image, listed count, friends count, followers count, statuses count, screenname, description. Medium importance (score: 0.3-0.6): geo-enabled, protected status, verified status, updated timestamp, created at timestamp, URL, language. Low importance (score: 0.0-0.3) location.

However, to facilitate fair comparison and aggregation of feature importance across models, it is essential to normalize the obtained importance scores. Normalizing the importance scores obtained from each model to ensure consistency in scale across all features. Techniques like min-max scaling can be applied to rescale the scores to a common range, facilitating fair comparison and aggregation of feature importance across models. Min-max scaling is a technique used to normalize data by transforming it to a common scale. This method rescales the values of a feature to a specified range, typically between 0 and 1. The process involves subtracting the minimum value of the feature from each data point and then dividing it by the difference between the maximum and minimum values. As a result, the importance scores obtained from different models are adjusted to ensure consistency in scale across all features. This normalization facilitates fair comparison and aggregation of feature importance across models, making it easier to interpret and analyze the relative importance of each feature [14].

– Facebook fake/real dataset

Following data preprocessing, the dataset is refined to include essential features crucial for subsequent analysis. These features encompass various aspects such as user attributes and profile characteristics, including

'id,' 'name,' 'screen\_name,' 'statuses\_count,' 'followers\_count,' 'friends\_count,' 'favourites\_count,' 'listed\_count,' 'created\_at,' 'URL,' 'lang,' 'location,' 'default\_profile,' 'profile\_image\_url,' 'profile\_banner\_url,' 'profile\_use\_background\_image,' 'profile\_background\_image\_url\_https,' 'profile\_text\_color,' 'profile\_image\_url\_https,' 'profile\_sidebar\_border\_color,' 'profile\_sidebar\_fill\_color,' 'profile\_background\_image\_url,' 'profile\_background\_color,' 'profile\_link\_color,' 'description,' 'updated,' and 'dataset.' From this comprehensive pool of features, the top twenty are carefully selected based on their significance for subsequent analysis. These include 'status count,' 'followers count,' 'friends count,' 'favorites count,' 'listed count,' 'created at,' 'URL,' 'lang,' 'location,' 'default profile,' 'description,' 'profile image URL,' 'profile banner URL,' 'profile use a background image,' 'profile background image URL https,' 'profile text color,' 'profile image URL https,' 'profile sidebar border color,' 'profile sidebar fill color,' and 'profile link color.' The selection process employs the k best method technique, utilizing the ANOVA F-statistic (f\_classif) scoring function. This technique measures the ratio of variances between classes to the variance within classes, making it suitable for assessing the statistical significance of numerical features concerning a categorical target variable. Features with higher F-statistic scores are prioritized as they offer greater informativeness for classification tasks.

### 2.3. Classification approaches

Two distinct approaches have emerged as prominent solutions in tackling complex classification tasks: machine learning approach and federated learning approach. Machine learning methods [15] encompass a wide spectrum of established techniques such as decision trees (DT), RF, SVM, LR, and ANN. These methods have demonstrated their prowess in diverse domains, leveraging historical data patterns to make predictions or classify data into predefined categories. In contrast, federated learning is a more recent paradigm with a privacy-centric and decentralized approach. It enables model training across multiple distributed devices or servers without sharing raw data. Instead, only model updates are exchanged, preserving data privacy. These two approaches offer unique advantages and trade-offs, allowing practitioners to choose the most suitable method based on their specific use cases, data availability, and privacy requirements.

#### 2.3.1. Machine learning

In the pursuit of achieving superior predictive performance, a comprehensive strategy often involves employing a suite of traditional machine learning models, including DT that partition the feature space based on decision rules, making them easy to interpret. They can handle both numerical and categorical data and are robust to outliers, LR models the probability of a binary outcome and provides interpretable results in terms of odds ratios. It is efficient for binary classification tasks and works well with large datasets, RF are an ensemble of DT that are robust to overfitting. They handle high-dimensional data well and provide feature importance scores, making them useful for feature selection, ANN inspired by the human brain, ANN capture complex relationships in data. They are effective for large-scale problems.

Naive Bayes (NB) is a probabilistic classifier based on the assumption of conditional independence between features. It is simple, performs well with small training data, and is particularly effective in text classification tasks, SVM that find the hyperplane that best separates classes in the feature space. They are effective in high-dimensional spaces, work with linear and non-linear data, and are robust to overfitting, and GB builds an ensemble of sequentially trained DT, leading to highly accurate predictions. It handles heterogeneous data well and can be used for both classification and regression tasks.

The process extends beyond mere model selection as it delves into ensemble learning. This approach entails choosing the most effective combination of models, such as LR, RF, SVM, GB, and ANN, and using stacking as the ensemble technique. When combined in a stacking ensemble, these models offer a diverse set of strengths that can collectively contribute to superior predictive performance. LR provides interpretability, RF and SVM offer robustness and feature handling capabilities, GB provides accuracy through sequential tree boosting, and ANN adds the capability to capture complex relationships. By leveraging the strengths of each model, the stacking ensemble can mitigate individual weaknesses and achieve enhanced predictive power.

Each base model is individually trained on the training data in stacking, and their predictions are collected for the validation dataset. These model predictions, along with the original features and target variable, are fed into a meta-learner, typically a gradient-boosting model. The goal is to harness the diverse strengths of each base model to enhance predictive accuracy and generalization, resulting in a robust and powerful ensemble approach that often outperforms the individual models [16], [17].

Aggregation techniques:

Aggregation techniques are methods used to combine the predictions of multiple base models to enhance predictive accuracy. One common type of aggregation technique is simple averaging, where the predictions of all base models are averaged to obtain the final prediction. Weighted averaging is another approach, where each base model's prediction is weighted differently before averaging, allowing for more influence from certain models. However, a more sophisticated aggregation technique known as stacking the

proposed ensemble method in our work has gained popularity in recent years [18]. Stacking involves training multiple base models and then combining their predictions using a meta-model. Unlike simple averaging, stacking introduces an additional layer of learning where the meta-model learns to combine the predictions of the base models. This approach allows for leveraging the diverse perspectives provided by the base models, which may use different algorithms or feature subsets. By capitalizing on the complementary strengths of individual models, stacking has the potential to significantly improve predictive accuracy. However, it is important to note that stacking may require additional computational resources and careful tuning of meta-parameters to achieve optimal performance.

### 2.3.2. Federated learning

In the context of tasks involving user personal data, federated learning has emerged as a groundbreaking approach that marries the power of machine learning with data privacy preservation. Federated learning allows for the collaborative training of machine learning models across multiple devices or servers while keeping sensitive user data decentralized and secure. In such tasks, where privacy and data protection are paramount concerns, federated learning shines by ensuring that user personal data remains on the local devices or servers, never leaving the user's control. Instead of transmitting raw data to a central server for model training, only model updates are shared, greatly reducing the risk of data breaches or privacy violations. This innovative technique allows for developing highly personalized and effective models without compromising the confidentiality of user information. Federated learning thus represents a significant step forward in addressing privacy and security challenges while utilizing machine learning to deliver customized and accurate results in tasks reliant on users' personal data. Dividing data into different virtual clients depends on the size of the data. In Instagram datasets, we used two clients only because of the small data size, while in Twitter and Facebook datasets, we used three clients, which refer to three subsets of data. The structure of federated learning is shown in Figure 2 [19], [20].

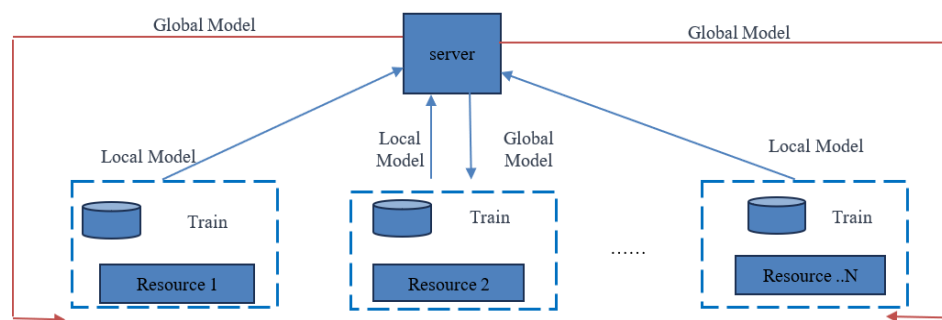


Figure 2. Federated learning structure

## 3. RESULTS AND DISCUSSION

This study delved into the impact of proposed approaches on detecting fake profiles across multiple platforms, while acknowledging that most of previous studies primarily focused on assessing methods on individual platforms and often overlooked data privacy concerns in this realm. The experiments entailed analyzing the performance of both machine learning and federated learning approaches across four datasets from three platforms. Seven machine learning algorithms were rigorously evaluated, including DT, LR, RF, SVM, NB, GB, and ANN. Performance metrics such as accuracy, precision, recall, and F1-score were utilized to gauge predictive capabilities on platform-specific datasets, revealing individual strengths and weaknesses. Subsequently, the stacking model, an ensemble technique combining base model predictions with original features and target variables, was introduced and its impact on overall performance discussed. The discussion delved into nuances within each dataset, showcasing where specific algorithms excelled and how stacking enhanced predictive capabilities. The stacking model aggregates predictions from the five most accurate base machine learning models, leveraging the complementary strengths of individual models to potentially enhance performance compared to any single model.

We find that although there are several ensemble methods exist in machine learning, such as bagging and boosting, but stacking has advantages over these methods. These include model flexibility, improved performance, and the ability to capture complex data relationships in a non-linear manner. Furthermore, the selection of proficient base learners significantly influences the overall performance of the stacked model, prompting extensive experimentation in this phase. Additionally, federated learning emerged

as a promising approach, particularly in scenarios involving personal and sensitive data, as it preserves data privacy while still achieving commendable results. However, it is worth noting that federated learning typically requires a higher execution time compared to traditional machine learning algorithms.

Our experiments findings underscore the efficacy of the stacking method when leveraging effective base learners from machine learning algorithms in tackling classification challenges, as demonstrated in our study focused on identifying fake accounts across various social platforms using four distinct datasets. Overcoming diverse challenges, ranging from data bias to feature selection, valuable lessons have been learned from the mistakes encountered. Biases in training data and incomplete feature engineering can significantly impact model performance. To address biases, researchers must be vigilant and mitigate them during preprocessing. Additionally, prioritizing comprehensive feature engineering is crucial to capture relevant information effectively. Implementing corrective measures based on these lessons aims to enhance model robustness and reliability in real-world scenarios. This approach is exactly what we adopt in our study.

### 3.1. Evaluation metrics

We present the metrics used to evaluate the results to select the best-supervised machine learning algorithm. We show the model accuracy, precision, recall, and F\_Score, which are calculated as [21], [22]:

- Accuracy: accuracy measures the proportion of correctly predicted instances out of the total number of instances in the dataset as described in (1).

$$\frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Where, true positives (TP) are the number of correctly predicted positive instances, true negatives (TN) are the number of correctly predicted negative instances, false positives (FP) are the number of instances incorrectly predicted as positive, and false negatives (FN) is the number of instances incorrectly predicted as negative.

- Precision: precision quantifies the accuracy of positive predictions calculated as described in (2).

$$\frac{TP}{TP+F} \quad (2)$$

- Recall (sensitivity): recall measures the model's ability to identify all positive instances correctly. calculated as described in (3).

$$\frac{TP}{TP+FN} \quad (3)$$

- F-score (F1\_Score): the F1 score is the harmonic mean of precision and recall and provides a balanced measure of the model's performance as shown in (4).

$$2 * \frac{\text{precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

These formulas offer precise calculations for each metric, facilitating a comprehensive evaluation of the model's performance in classification tasks.

### 3.2. Comparative analysis of results

The experiments with Instagram datasets are depicted as follows: Table 1 illustrates the machine learning results obtained for Dataset 1, while Table 2 showcases the outcomes for Dataset 2. It is noteworthy that the stacking model which combines predictions from the best five base models, outperforms each individual model. This enhancement in performance highlights the effectiveness of the stacking approach in leveraging the collective strengths of diverse models to achieve superior predictive accuracy.

Table 3 displays the results of our proposed model, which surpasses the approach outlined in reference [6]. The authors of that study achieved an accuracy of 92.9% by employing the RF algorithm and oversampling the data using the SMOTE-NC algorithm. Additionally, our approach outperforms the results reported in [8], where an accuracy of 82% was achieved using the KNN algorithm on Dataset 1. Table 4 presents a comparative analysis using Dataset 2, wherein our approach is juxtaposed with the method outlined in reference [23]. The reference study utilized various machine learning algorithms and achieved an 86% F-score using SVM and NN. However, our proposed model attained superior performance in comparison.



Table 1. Dataset 1 results using machine learning

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F-score (%)
SVM	92.4	90.3	91	88.3
RF	92.9	89.2	88.1	89.2
GB	93	90	90	92
LR	91.2	87.3	89	88
NB	89.1	85.2	86.2	86.1
DT	89.8	86.3	82.3	88.1
ANN	91	89	88	87
Stacking Model	94.8	93.2	92.1	94.2

Table 2. Dataset 2 results using machine learning

Classifier	Accuracy (%)	Precision (%)	Recall (%)	Fscore (%)
SVM	91.3	91.3	82	86.3
RF	92.2	88.2	79.9	84.2
GB	92.8	90.3	91.2	88.3
LR	83.4	80.3	70	75
NB	86.2	85.2	68.2	78.1
DT	89.3	86.3	82.3	88.1
ANN	90.4	89	84	86
Stacking Model	93.5	93.2	87.1	89.2

Table 3. Comparing proposed model results with [6], [8] on Dataset 1

Ref	Model	Accuracy (%)	Precision (%)	Recall (%)	F-score (%)
[6]	RF with SMOTE-NC	92.9	-	-	-
[8]	KNN	82	-	-	-
	Proposed Model	94.8%	94	96	96

Table 4. Comparing proposed model results with [23] on Dataset 2

Ref	Model	F-score (%)
[23]	SVM, NB, NN, LR	86
This work	Proposed Model	89

Table 5 in our study presents a comparison of results obtained from different machine learning algorithms alongside the performance of our stacking model. This table offers valuable insights into the efficacy of ensemble techniques in enhancing predictive accuracy across various datasets and scenarios. Furthermore, Table 6 provides an analysis of our proposed model's performance compared to the methodologies described in references [24], [25]. Specifically focusing on X (Twitter) Dataset 3, this comparison highlights the competitive advantage of our model in addressing the unique challenges posed by this dataset. Consequently, it contributes significantly to the broader comprehension of effective machine-learning strategies for the given task.

Table 5. Dataset 3 results using machine learning

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F-score (%)
SVM	91.3	91.3	82	86.3
RF	98.2	88.2	79.9	84.2
GB	98.8	90.3	91.2	88.3
LR	92.4	80.3	70	75
NB	94.2	85.2	68.2	78.1
DT	91.3	86.3	82.3	88.1
ANN	96.4	89	84	86
Stacking model	99.4	92.2	87.1	89.2

Table 6. Comparing proposed model results with [24], [25] on Dataset 3

Ref	Model	Accuracy (%)
[24]	RF	98
[25]	SVM	95.7
This work	Proposed model	99.4

Table 7 presents the performance results of various machine learning algorithms and the stacking model when applied to face book data (Dataset 4). This table provides valuable insights into the effectiveness of these approaches in tackling the unique characteristics of Dataset 4. Table 8 compares our proposed model and the model presented in [26]. This comparison is specifically carried out on Dataset 4, shedding light on each approach's relative strengths and weaknesses within the context of this dataset.

Table 7. Results using machine learning on Dataset 4

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F-score (%)
SVM	99	89.5	92.3	92.3
RF	99.6	92.5	93.2	95.3
GB	87.3	88.3	85.4	88.4
LR	95.1	90.2	85.5	91.5
NB	90.4	87.5	89.4	90.3
DT	80.2	74.7	79.5	80
ANN	93.4	78.3	80.2	76.4
Stacking model	99.3	93.2	95.4	95.5

Table 8. Comparing the proposed model with [26] on Dataset 4

Ref	Model	Accuracy (%)
[26]	NB, j48 and RF	99.6
This work	Proposed Model	99.8

In the comparative analysis of federated learning and the stacking model of machine learning algorithms, we divided data from Instagram, Twitter, and Facebook into virtual resources to accommodate different dataset sizes. Table 9 evaluated the two approaches based on execution time and accuracy measures. Notably, federated learning consumed more time than the stacking model due to several factors, including the need for data aggregation across resources, communication overhead for model updates, and complexities associated with federated optimization. This observation highlights the trade-offs between data privacy, the strength of federated learning, and its computationally intensive nature. The choice between the two approaches should be based on the specific use case available computational resources and the need for security.

Table 9. A comparison between federated learning versus stacking model

Dataset	Federated learning Execution time	Federated learning_ accuracy (%)	Stacking model Execution time (s)	Stacking_model_ accuracy (%)
Dataset 1(2 resources)	1.3s	96	0.8	94.8
Dataset 2(2 resources)	1.2s	95	0.7	93.5
Dataset 3(3 resources)	2.3s	98	1.4	99.4
Dataset 4(3 resources)	2.1s	99	1.5	99.8

Addressing limitations and their potential impact on the results is crucial for ensuring the credibility and reliability of the study findings. While the recommended approaches indeed outperform other methods, it's crucial to acknowledge the limitations inherent in the model. One such limitation pertains to generalizability. While our study demonstrates the efficacy of the stacking model in identifying fake profiles across various social media platforms, it is essential to recognize potential limitations in its generalizability to other domains. Different domains possess distinct characteristics and challenges that might affect the stacking model's performance differently. Moreover, the performance of machine learning models, including the stacking model, is highly dependent on the quality and diversity of the training data. Our study utilized datasets specific to Instagram, Twitter, and Facebook, which may not fully capture the diversity of fake profiles across all social media platforms such as LinkedIn, Snapchat, TikTok, Reddit, and Pinterest. Additionally, biases present in the training data could lead to biased predictions, particularly in real-world applications where data may be incomplete or skewed. Addressing these concerns is crucial for ensuring the reliability and applicability of the stacking model in broader contexts.

For future research, further refinement of the stacking model warrants exploration, which could involve investigating various ensemble techniques and meta-model architectures to potentially elevate predictive performance to even greater heights. Moreover, incorporating advanced feature engineering methods and integrating additional data sources may yield valuable insights, enhancing the model's robustness. Additionally, subjecting the developed framework to real-time data streams and assessing its performance in dynamic environments could provide invaluable insights into its practical viability and scalability. Lastly,

extending the application of federated learning to diverse domains and adapting it to different data types could broaden its utility and foster advancements in privacy-preserving machine learning methodologies.

In conclusion, the choice between federated learning and traditional machine learning approaches hinges on the specific requirements of the task at hand. If prioritizing data privacy and execution time is not a concern, federated learning presents an ideal solution. Conversely, for tasks where data privacy is not a primary consideration and swift, accurate results are imperative, traditional machine learning approaches offer a viable alternative.

#### 4. CONCLUSION

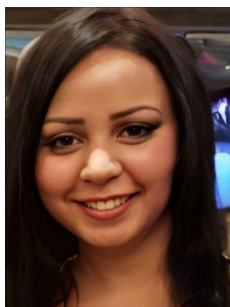
The research addresses fake profile detection on major social media platforms: Instagram, X (Twitter), and Facebook. We compared a stacking model of machine learning algorithms with federated learning. The stacking model achieved high accuracies: 96% and 95% on Instagram datasets, 99.4% on the X dataset, and 99.8% on the Facebook dataset. Federated learning offers data privacy benefits but performs slightly lower than the stacking model, highlighting a trade-off between performance and privacy. Our study introduces an effective stacking model, emphasizes ethical data privacy considerations, and underscores the model's adaptability and superior performance across various datasets. Future research could optimize the stacking model, integrate advanced feature engineering, apply the framework to real-time data, and explore federated learning in other domains. Future research can optimize the stacking model with different ensemble techniques and integrate advanced feature engineering and additional data sources to improve robustness. Additionally, applying the framework to real-time data and exploring federated learning in other domains could enhance practical utility and advance privacy-preserving machine learning.




#### REFERENCES

- [1] N. G. Kerrysa and I. Q. Utami, "Fake account detection in social media using machine learning methods: literature review," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 6, pp. 3790–3797, Dec. 2023, doi: 10.11591/eei.v12i6.5334.
- [2] B. Goyal *et al.*, "Detection of fake accounts on social media using multimodal data with deep learning," *IEEE Transactions on Computational Social Systems*, pp. 1–12, 2024, doi: 10.1109/TCSS.2023.3296837.
- [3] M. Taha, H. H. Zayed, M. Azer, and M. Gadallah, "Automated COVID-19 misinformation checking system using encoder representation with deep learning models," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 12, no. 1, Mar. 2023, doi: 10.11591/ijai.v12.i1.pp488-495.
- [4] L. Bioglio and R. G. Pensa, "Analysis and classification of privacy-sensitive content in social media posts," *EPJ Data Science*, vol. 11, no. 1, Dec. 2022, doi: 10.1140/epjds/s13688-022-00324-y.
- [5] Y. Chen, L. Liang and W. Gao, "FedADSN: Anomaly detection for social networks under decentralized federated learning," in *2022 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI)*, Dalian, China, 2022, pp. 111–117, doi: 10.1109/CCCI55352.2022.9926623.
- [6] F. A. J. Sarhan and E. A. Mattar, "Fake accounts detection in online social networks using hybrid machine learning models," *International Journal of Simulation: Systems, Science & Technology*, vol. 24, no. 2, pp. 2.1–2.5, 2023, doi: 10.5013/ijssst.a.24.02.02.
- [7] R. R. Rostami, "Detecting fake accounts on twitter social network using multi-objective hybrid feature selection approach," *Webology*, vol. 17, no. 1, pp. 1–18, May 2020, doi: 10.14704/WEB/V17I1/a204.
- [8] A. N. Hakimi *et al.*, "Identifying fake account in facebook using machine learning," in *Advances in Visual Informatics: 6th International Visual Informatics Conference, IVIC 2019, Bangi, Malaysia, November 19--21, 2019, Proceedings 6*, 2019, pp. 441–450, doi: 10.1007/978-3-030-34032-2\_39.
- [9] M. Mukherjee and M. Khushi, "SMOTE-ENC: A novel SMOTE-based method to generate synthetic data for nominal and continuous features," *Applied System Innovation*, vol. 4, no. 1, Mar. 2021, doi: 10.3390/asi4010018.
- [10] T. Wongvorachan, S. He, and O. Bulut, "A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining," *Information*, vol. 14, no. 1, pp. 1–15, Jan. 2023, doi: 10.3390/info14010054.
- [11] A. A. Ali, "The role of feature engineering in enhancing machine learning models," *OSF Preprints*, doi: 10.31219/osf.io/vf3rg.
- [12] R. Rousseau, "Jaccard similarity leads to the Marczewski-Steinhaus topology for information retrieval," *Information Processing & Management*, vol. 34, no. 1, pp. 87–94, Jan. 1998, doi: 10.1016/S0306-4573(97)00067-8.
- [13] M. Büyükköçeci and M. C. Okur, "A comprehensive review of feature selection and feature selection stability in machine learning," *Gazi University Journal of Science*, vol. 36, no. 4, pp. 1506–1520, Dec. 2023, doi: 10.35378/gujs.993763.
- [14] H. Klaus and E. Blessing, "Normalization and Standardization: Methods to preprocess data to have consistent scales and distributions," *ResearchGate*, pp. 1–10, 2023.
- [15] J. Sen, *Machine learning: algorithms, models, and applications*, London, UK: IntechOpen, 2022.
- [16] S. U. Hassan, J. Ahamed, and K. Ahmad, "Analytics of machine learning-based algorithms for text classification," *Sustainable Operations and Computers*, vol. 3, pp. 238–248, 2022, doi: 10.1016/j.susoc.2022.03.001.
- [17] P. Chakraborty, M. M. Shazan, M. Nahid, M. K. Ahmed, and P. C. Talukder, "Fake profile detection using machine learning techniques," *Journal of Computer and Communications*, vol. 10, no. 10, pp. 74–87, 2022, doi: 10.4236/jcc.2022.1010006.
- [18] T. Zhou and H. Jiao, "Exploration of the stacking ensemble machine learning algorithm for cheating detection in large-scale assessment," *Educational and Psychological Measurement*, vol. 83, no. 4, pp. 831–854, Aug. 2023, doi: 10.1177/00131644221117193.
- [19] P. M. Mammen, "Federated learning: opportunities and challenges," *arXiv-Computer Science*, pp. 1–5, Jan. 2021.
- [20] C. Zhang, Y. Xie, H. Bai, B. Yu, W. Li, and Y. Gao, "A survey on federated learning," *Knowledge-Based Systems*, vol. 216, Mar. 2021, doi: 10.1016/j.knosys.2021.106775.




- [21] R. Yacoub and D. Axman, "Probabilistic extension of precision, recall, and F1 score for more thorough evaluation of classification models," in *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 79–91. doi: 10.18653/v1/2020.eval4nlp-1.9.
- [22] Ž. Đ. Vujovic, "Classification model evaluation metrics," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, 2021, doi: 10.14569/IJACSA.2021.0120670.
- [23] F. C. Akyon and M. Esat Kalfaoglu, "Instagram fake and automated account detection," in *2019 Innovations in Intelligent Systems and Applications Conference (ASYU)*, Izmir, Turkey, IEEE, Oct. 2019, pp. 1–7. doi: 10.1109/ASYU48272.2019.8946437.
- [24] A. Homs, J. Al Nemri, N. Naimat, H. A. Kareem, M. Al-Fayoumi, and M. A. Snober, "Detecting Twitter fake accounts using machine learning and data reduction techniques," in *DATA*, 2022, pp. 88–95. doi: 10.5220/0010604300002993.
- [25] D. R. Patil, T. M. Pattewar, V. D. Punjabi, and S. M. Pardeshi, "Detecting fake social media profiles using the majority voting approach," *EAI Endorsed Transactions on Scalable Information Systems*, vol. 11, no. 3, pp. 1-18, Feb. 2024, doi: 10.4108/eetis.4264.
- [26] Y. Elyusufi, Z. Elyusufi, and M. A. Kbir, "Social networks fake profiles detection using machine learning algorithms," in *Innovations in Smart Cities Applications Edition 3*, 2020, pp. 30–40. doi: 10.1007/978-3-030-37629-1\_3.

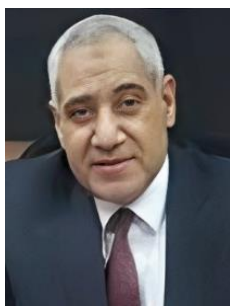
## BIOGRAPHIES OF AUTHORS






**Marina Azer**    is an artificial intelligence specialist, she was a Lecturer Assistant in Faculty of Computer Science, Modern Academy, Egypt, since 2012. She received B.Sc. in computer science in 2012 at Faculty of Computer Science and M.Sc. in Computer Science in 2017 at Faculty of Computer Science, Helwan university. She has worked on several research topics, her research interests are neural network, machine learning, deep learning, and pretrained models. She can be contacted at email: marina.essam991@gmail.com.






**Hala H. Zayed**    received the B.Sc. degree (Hons.) in electrical engineering and the M.Sc. and Ph.D. degrees in electronics engineering from Benha University, in 1985, 1989, and 1995, respectively. She is a professor at Faculty of Engineering, Egypt University of Informatics (EUI), Cairo, Egypt. Her research interests include computer vision, biometrics, machine learning, image forensics, and image processing. She can be contacted at email: hala.zayed@eui.edu.eg.



**Mahmoud E. A. Gadallah**    received the B.Sc. in Electrical Engineering in 1979, the M.Sc. in 1984 Faculty of Engineering, Cairo University, and Ph.D. in 1991 from Cranfield Institute of Technology (Cranfield University now), United Kingdom. He is now a professor at the modern academy for computer science and Management Technology, Cairo, Egypt. He has worked on several research topics as image processing, pattern recognition, computer vision, and natural language processing. He can be contacted at email: mgadalla1955@gmail.com.



**Mohamed Taha**    is an Associate Professor at Department of Computer Science, Faculty of Computers and Artificial intelligence, Benha University, Egypt. He received his M.Sc. degree and his Ph.D. degree in computer science at Ain Shams University, Egypt, in February 2009 and July 2015. He is the founder and coordinator of "networking and mobile technologies" program, Faculty of Computers and Artificial Intelligence, Benha University. His research interest's concern: computer vision (object tracking-video surveillance systems), digital forensics (image forgery detection-document forgery detection-fake currency detection), image processing (OCR), computer network (routing protocols-security), augmented reality, cloud computing, and data mining (association rules mining-knowledge discovery). He has contributed more than 30+ technical papers in international journals and conferenceshe can be contacted at email: mohamed.taha@fci.bu.edu.eg.