

## EXPLORING VARIOUS FEATURES FOR IDENTIFICATION OF FAKE PROFILES USING MACHINE LEARNING

Usama Hafeez<sup>\*1</sup>, Yasir Ghayor Malik<sup>2</sup>, Muhammad Waqas Haider<sup>3</sup>, Muhammad Imran Hafeez<sup>4</sup>,  
Anam Zahoor<sup>5</sup>, Rashid Mehmood Gondal<sup>6</sup>

<sup>\*1,2,3,4,5</sup>Department of Computer Science & IT, University of Lahore, 40100, Punjab, Pakistan

<sup>6</sup>Department of Computer Science, University of Sargodha, 40100, Punjab, Pakistan

<sup>1</sup>mhmd.usama.hafeez@gmail.com, <sup>2</sup>yasir.ghayor@cs.uol.edu.pk, <sup>3</sup>whaider491@gmail.com,

<sup>4</sup>muhammad.hafeez@cs.uol.edu.pk, <sup>5</sup>malikanam06@gmail.com, <sup>6</sup>rashidmehmoodit@gmail.com

DOI: <https://doi.org/10.5281/zenodo.23160936>

### Keywords

Fake profile, Instagram, Classification algorithm, Fake followers, Deep learning, Ensemble learning, Machine learning, Social Media.

### Article History

**Received:** 25 June 2026

**Accepted:** 16 September 2026

**Published:** 02 October 2026

**Copyright** @Author

**Corresponding Author:** \*

Usama Hafeez

### Abstract

Instagram is gaining popularity among youth as it allows for self-expressiveness, promotion, and social connection. However, this popularity raises an issue of fake profile creation being used of disinformation, impersonation, and fraud, critical for user trust and the integrity of social networking platforms. In this research study, aim is to identify the fake Instagram profiles and classify them either fake or real. Focusing on extracting a comprehensive set of dataset features, such as user profile information, tagging and hashtags, content characteristics, interaction metrics, and content similarity measures, using the Fake Authentic Instagram User Dataset. Using the feature selection techniques, such as relief, correlation, gain ratio, and information gain, for ranking top features. For empirical analysis, a diverse range of conventional machine learning and deep learning techniques applied to evaluate their efficacy in classifying profiles. The algorithms that undergo the evaluation consist of shallow models, ensemble models deep learning models. The performance of each algorithm is estimated by standard performance evaluation measures of classification. The results show that Random Forest, an ensemble-learning model, outperforms other models with an accuracy of 95.10% in distinguishing between the fake and real Instagram users' profile.

## INTRODUCTION

In today's digital world, social media platforms are becoming an important part of our daily life. With the increasing popularity of platforms such as Twitter, Facebook, Instagram, and linkedin people are investing more in their brands promotion for greater revenue and social connections. Instagram, the popular social network gains its popularity among different age groups due to ease of expressiveness, promoting brands and influencers, and socially interacting

with followers and followings via post, stories and Instagram live. To destroy the user trust and credibility of social media platforms fake profiles are being created for impersonation, i.e., to pretend to be someone else. Spam, disinformation, or to artificially inflate metrics like follower counts and likes, aim to misrepresent the actual users are the unethical reasons behind the creating of these fake profiles. Instagram and other social networks profiles include username, bio, and profile picture, does

not validate based on user identity. People with unethical intentions can take advantage of such social network loopholes for malicious activities [1]. Critical for integrity and privacy of Instagram users and other social media platforms. For digital marketers and online communities fake user profiles are becoming the major problem. [2]. In this research study, identification of fake Instagram profiles and then classify these profiles as either fake or real Instagram profiles.

In order to protect the personal information and preserving the integrity of the online communications in our connected world, first theirs need to understand the nature of fake profile. Click farm profiles produce fake friends, likes, and views and catfish accounts use stolen or made-up personal information to pretend to be other people. Sock puppet profiles are control by one person and created by combining number of fake profiles [3]. In 2022, 58.19% percent of all mega-influencers were involved in fraudulent operations. On average, 49.32% of influencers participated in fraudulent actions that year [4]. The anticipated monthly active user count for Instagram was about 2.3 billion people in June 2023 approximately 4.75% of all monthly active users fall into this category [5].

When it comes to detecting the fake profiles on social media networks like Instagram, machine learning offers smart and effective solutions. Analyzing the immense datasets for abnormalities and inconsistencies that can indicate the fake users profile is an integral aspect of these methods. Deep learning models, ensemble models, and shallow models are among the most popular way for detecting fake Instagram profiles. Profile maturity, ways of interaction, and posted contents are quickly being detected by Shallow ML models. Combining multiple ensemble techniques can have improved the accuracy and precisions for detection. The more innovative fake Instagram profiles having complex interaction patterns and user behavior are detected using the Deep Learning algorithms [6]. The ever-changing nature for the creation of fake accounts, these classification algorithms needs to

adapt changes in order to maintain their effectiveness.

In this research study, the goal is to inspect how different dataset features, categorized by their importance in feature selection techniques such as correlation, information gain, relief, and gain ratio for the detection of fake user profiles affects the efficiency of different machine learning algorithms. Systematically comparing the outputs of machine learning models to formulate the performances across various dataset features, features categories and their combinations. To identify optimal technique, the model evaluates using various performance evaluation measures such as accuracy, precision, recall, and F1-score. This research will not only make significant contributions to the theoretical breakthroughs in the field of data science, but it will also provide valuable practical insights for improving the process of feature engineering in predictive modeling.

This research study achieves the following contributions:

- Carrying out feature selection to find top k features using diverse techniques such as Information Gain, Gain Ratio, Correlation, and Relief-F.
- Application of diverse ML shallow models such as SVM, KNN, Decision Trees, & Naive Bayes and ensemble learning algorithms like adaboost, xgboost, & Random Forests, and different DL models such as CNN, RNN, and LSTM networks for the prediction of fake Instagram profile.
- Analysis of different algorithms and dataset feature categories shows that Random Forest achieves highest accuracy of 95.10% and user profile information and interaction metrics are the most predictive features.
- Comparison of the proposed methods with existing studies using performance evaluation measures using standard metrics like accuracy, precision, recall, and F1-Score.

The remaining sections are organized as follows. Section 2 presents the related work of existing studies, while Section 3 provides an in-depth discussion of the research methodology and setup. Section 4 deals with the experimental

analysis and results, and Section 5 gives the overall conclusion for the detection of fake profile.

## Related Work

Setting up fake profile involves creating an account on a social media platform under a false name or with falsified information about a fictional character or person. Such accounts frequently participate by following a large number of people and submitting photos that are either fake or created by artificial intelligence. Detecting and removing these profiles to protect user privacy, as they facilitate phishing, identity theft, and malware distribution. Their removal prevents disinformation and impersonation of brands and public figures on social media. Using machine-learning algorithms for detection, which scrutinize engagement patterns, account creation speed, and content validity. Network analysis reveals interconnected networks of fake profiles, whereas behavioral analysis reviews posting habits and interactions.

Identification of fake profiles on Twitter and other social media platforms is the primary objective of this research. These fake accounts identified by using the technique of network analysis and similarity measurements, which are determined by adjacency matrices. To find out how well the Medium Gaussian SVM worked, Mohammadrezaei *et al* [7] applied it to twitter dataset.

A way to identify fake twitter accounts is proposed by Rostami and Karbasi [8] is the minimum redundancy-maximum relevance method which is the basically a multi objective hybrid feature selection technique. This technique detects features that are both distinguishing and effective, which makes the classification process work better. The two twitter datasets are utilized for applying and evaluating the approach using Random Forest, Naive Bayes, and Support Vector Machine classifiers. With a smaller but more valuable feature set, the concept outperformed previous approaches in terms of accuracy and reliability.

There still is an on growing issue of spam and increasing disinformation fake accounts are being creates for this purpose. Kaushik *et al* [9] have proposed an advanced structure like neural network and along with it incorporating SVM and Naive Bayes algorithms for detection. These algorithms shows significant performance with an accuracy of 83.5%. Offering the cybersecurity, the method for overcoming the serious threat of Fake profile for the online social networking sites.

Detection of fake Twitter profiles Gangan *et al* [10] can addressed the problem by incorporating the ensemble learning techniques with the Naive Bayes and Decision Tree algorithms. This ensemstack technique achieves an accuracy rate of about 93.93% for the detection of fake profiles. The reliability and security of were Twitter increased by this ensemstack technique, which minimizes the risk related of a fake profile.

To improve the reliability and privacy of social networking sites especially the fake Instagram profiles Sheikhi *et al* [11] proposed the Bagging classifier technique. This is Decision Tree based machine learning approach capable of identifying the fake profiles by minimizing the prediction variance and improving the classifiers accuracy. Approximately 98% of accuracy achieved using Bagging classifier indicating its effectiveness when detecting the fake accounts.

Dey *et al* [12] resolved the issue of fake profile by incorporating the use of Random Forest and Logistic Regression to distinguish between the fake and real Instagram accounts comprising of unique profile features. Reviewing different user's profiles for their classification the model exhibits the notable performance evaluation measures especially in terms of accuracy 90.8 % and 92.5% for both Logistic Regression and Random forest respectively.

Combining different ensemble learning methods like Bagging, Stacking, Boosting, Blending, and Voting. Varol Arisoy and Tunc Abubakar [13] identifies the fake social networking profile by comparing the performance of the techniques. After applying the techniques Ada Boost shows and exceptional performance of about 91.92% accuracy as compared to the Bagging classifiers

having an accuracy of 90.41%. Their study shows that the ensemble techniques show greater effectiveness when detecting fake user profiles.

Nail *et al* [14] shows that using different ensemble algorithms like catboost, adaboost and Extra Tree Classifier for detection of fake profiles. Using the Instagram dataset in order to classify the user's profile that are fake. Their results show that about 96% of accuracy achieved for the catboost ensemble method showing its effectiveness in the process of detection. This highlights the greater concern that there is a need of continuous development in machine learning techniques in order to adapt the ever-changing nature of new ways for fake profile creation.

Lubis *et al* [15] study aims at identifying fake twitter profiles by evaluating the effectiveness of deep neural networks together with transfer learning techniques, highlighting the drawback of conventional detection technique. Exhibiting the impressive increase in the model accuracy of about 92.9% in terms of transfer learning and 86% in terms of deep learning technique. Their study highlights the reliability and capability of transfer learning method for the detection of fake user identities on social networking site like twitter and Instagram.

Analysis of twitter tweets content characteristics for determining the user account is either real or fake. For this purpose, the Bidirectional Gated Recurrent Unit model is applied. By ensuring that the syntactic and semantic context of the tweets are maintained glove terms are applied on the algorithm for the conversion of tweets into a vector space [16].

Implementing the application of using the sixteen profile and content based dataset features Amankeldin *et al* [17] developed a CNN framework for the detection of fake user accounts on social networking sites, particularly Instagram. Their study shows that the deep CNN algorithm can detect fake profile with an impressive accuracy of 99.4 %. Fake profile detection prevented using deep learning models, which may assist social media platforms in maintaining their dependability and integrity.

Detecting the fake user accounts on social media platform like Twitter Bhosale and Mane [18] proposed and innovative approach combining the machine and deep learning algorithms. Applying conventional models on twitter dataset like Naive Bayes, RF, Ada Boost, SVM, Gated Recurrent Unit, and LSTM, along with them a new hybrid model presented, which is the combination of GRU and LSTM. This hybrid model achieves an impressive accuracy of approximately 98.7% outperforming the others. This hybrid model improves internet security, privacy, and lead to cyber security.

Combining different ensemble learning methods like Bagging, Stacking, Boosting, Blending, and Voting. Varol Arisoy and Tunc Abubakar [19] identifies the fake social networking profile by comparing the performance of the techniques. After applying the techniques Ada Boost shows and exceptional performance of about 91.92% accuracy as compared to the Bagging classifiers having an accuracy of 90.41%. Their study shows that the ensemble techniques show greater effectiveness when detecting fake user profiles.

Zarei *et al.* [20] investigated Instagram impersonation and the detection of ingenuine content using deep learning. The authors collected 2.2K impersonator profiles, 10K posts, 68K comments, and 90K likes from political, sports, and musician communities. Using clustering, impersonators were categorized as bots or fans. A DNN combining BERT, CNN, LSTM, and profile metadata classified posts as bot-generated, fan-generated, or genuine. The proposed model achieved 86% accuracy, outperforming Random Forest and demonstrating effective automated impersonator-content detection.

By integrating the Neural network and Decision tree algorithm Yashwanth and Jaisharma [21] proposed a new approach called the Novel Integrated Neural Network (NINN) that improves that accuracy of detecting fake Instagram user accounts. This can ultimately lead to the understanding of how automated fake profiles can negatively affects the user security and privacy. With an accuracy rate of 90.91% for

NINN as compared to DT having 85.45% accuracy, the analysis shows that NINN outperforms the Decision Tree. This improvement in fake account detection shows the capability of integrating different methods to solve complicated digital problems.

Pan and Sharma [22] developed a machine learning approach for detecting fake Instagram profiles, evaluating Logistic Regression, Decision Tree, SVM, xgboost, and ensemble models. Features such as follower count, engagement rate, account age, bio and profile picture presence, and follower-following ratio were used, with SMOTE addressing class imbalance. Xgboost and the Stacked model performed best, achieving 93.17% test accuracy and 0.97 AUC-ROC, outperforming SVM, and Decision Tree in identifying fake accounts on Instagram.

Plikynas *et al.* [23] systematically reviewed machine learning and deep learning approaches for fake news, propaganda, and disinformation detection across authors/disseminators, textual content, and social impact. Using a PRISMA-based, they found that hybrid models integrating user, content, and social-network features provide more comprehensive detection than single-source approaches. The authors further proposed a foundational FNPDP model integrating diverse datasets and ML/DL techniques for adaptable detection and mitigation.

Fake Instagram profiles pose significant challenges by spreading misinformation and influencing public opinion, particularly during

critical events. The above related work reveals the lack of implementing a) feature selection techniques like Relief, Correlation, Information Gain, and Gain Ratio to identify the top 10 features in datasets; b) the analysis of classification outcomes for dataset; and c) the need for more distinct categorization of dataset and applying classification to individual features, category, and combined dataset categories. Applying machine-learning techniques such as Naive Bayes, Decision Trees, and KNN, ensemble learning methods such as adaboost, xgboost, and Random Forests, and deep learning models such as cnns, rnns, and lstms to obtain the classification results of the aforementioned points. The goal is to create an effective system for identifying and limiting the impact of fake profile.

## 1. Proposed Research Methodology

The fundamental elements of the research process, include the handling of datasets, preprocessing data, feature ranking techniques, application machine learning models, and the metrics for evaluating model performance. The Figure 1 highlights the key elements addressed, which include data preprocessing, as well as the utilization of diverse machine learning models and performance metrics. Outlining the importance of each step of the research process, with a focus on how essential it is to properly gather, organize and prepare data and analyze model to ensure the reliability of the results.

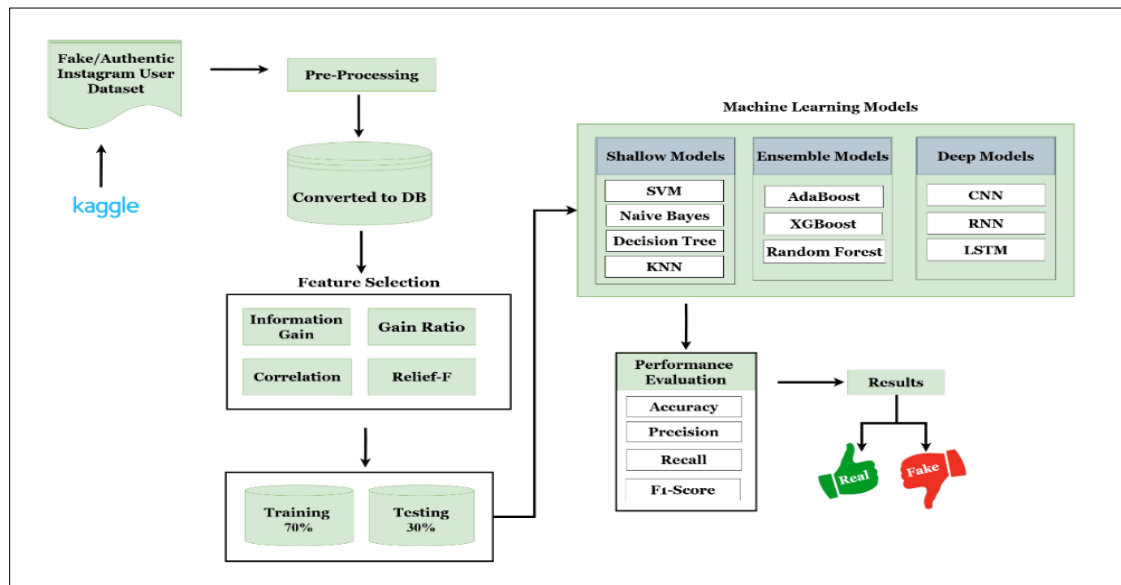


Figure 1. Framework Diagram for Research Methodology

### 1.1 Dataset Acquisition

For the identification and classification of Fake Instagram Profiles, using various machine-learning techniques Kaggle<sup>1</sup> dataset was utilized. For the purpose of efficient data retrieval and processing during analysis, the dataset was organized in the Comma Separated Values (CSV) format.

### 1.2 Data Preprocessing

The initial step of data analysis process is the preprocessing of the dataset. Before applying preprocessing and the application of machine-learning techniques, first need to clean and organize the unprocessed data. The dataset contains categorical labelling like ('r', 'f'). To convert the categorical labels to numerical values label encoding was applied by creating a label-mapping dictionary. The reason for converting these labels is to ensure that algorithms can interpret the data correctly, as numerical inputs are required for computations during model training, improving the reliability and accuracy of model predictions. Implementing preprocessing step will help you ensure that the dataset is well

suited for any further analysis or machine learning tasks, providing a robust foundation for building accurate and effective models.

### 1.3 Feature Engineering

The dataset feature used to train the classification model are outlined in figure 2. There are five separate categories into which these features fall, i.e., User Profile Information, Content Characteristics, Interaction Metrics, Tagging and Hashtags, and Content Similarity Measures. When it comes to identifying false profiles on social networking sites, the dataset features are quite dependent on one another. Information in a user's profile, like the number of posts, followers, and followings, can tell you a lot about how real and active a profile is. If an account has frequent posts & its likes and comments or followers compared to the number of followings, it might be a fake profile that being used for spamming or other illegal activities. Similarly, biography length, profile photo, and external links can separate real individuals from automated or fake profiles. Content characteristics, such as the average length of captions and the proportion of media types other than images, might show trends in scripted or

automated content creation, a feature that is frequently indicative of fake profiles. One popular approach employed by fake profile users is to artificially inflating a high number of

fake comments or likes. The interaction data, including the engagement rate for likes and comments, can help detect these account.

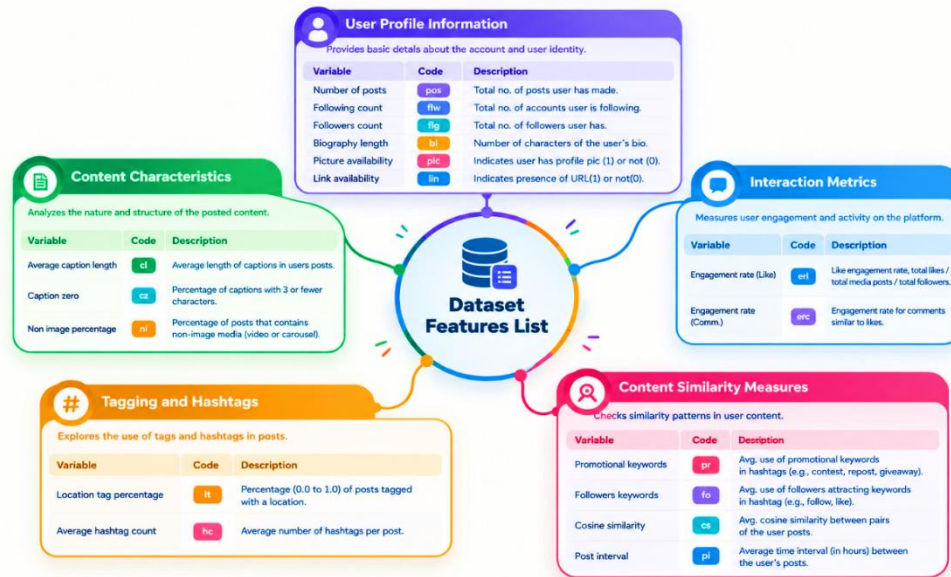


Figure 2. Dataset Features List

Fake profile networks often use coordinated attempts to promote specific material or attract followers. However, the tagging and hashtag functionalities, along with content similarity measurements, can identify these efforts.

#### 1.4 Feature Selection

Feature selection techniques are essential for determining the most significant features in a dataset. These methods help prioritize variables based on their importance in predictive modeling. The feature ranking techniques employed in this study include:

##### 1.4.1 Information Gain

A metric that quantifies the amount of information gained about the class variable when a specific feature is recognized. It helps in understanding the significance of features in predicting the target variable and plays a crucial role in feature selection and model building. Higher information gain values indicate features that are more informative for classification tasks.

The Equation 1 shows the Information Gain IG for a feature A with respect to the target variable T as:

$$IG(A, D) = Entropy(D) - \sum_{j=1}^K \left( \frac{|D_j|}{|D|} \right) Entropy(D_j)$$

Where  $|D|$  and  $A$  is the total dataset samples and attributes respectively.  $|D_j|$  is the number of samples in each subgroup after splitting. Entropy ( $D$ ) is the entropy of the target variable "class" before the split. The expression  $\sum_{j=1}^K \left( \frac{|D_j|}{|D|} \right) Entropy(D_j)$  is the weighted sum of the entropies for each label after the split.  $K$  is the number of labels  $A$  can take, and  $j$  indexes each label when calculating the sum.

##### 1.4.2 Gain Ratio

Gain Ratio is a measure used to evaluate how well a feature splits the data into classes while considering the intrinsic information content. It helps determine how well features partition data and their identifying strength. The Gain Ratio

computes by first calculating the Information Gain and then normalizing this value by the intrinsic information of the splitting attribute. To calculate the gain ratio of the dataset attributes we need to calculate the information gain and intrinsic value.

$$(1) IG(A, D) = Entropy(D) -$$

$$\sum_{j=1}^K \left( \frac{|D_j|}{|D|} \right) Entropy(D_j)$$

$$Gain\ Ratio(A, D) = \frac{IG(A, D)}{Intrinsic\ Info(A)}$$

### 1.4.3 Correlation

Correlation analysis conducted to understand the relationship between different features in the dataset. This analysis helps in identifying any interdependencies between variables and their impact on the target variable, aiding in feature selection and model performance improvement. The equation 5 shows how the Pearson correlation coefficient  $r$  calculated.

### 1.4.4 Relief

It can evaluate the relevance of features by considering the class labels of neighboring instances. By assessing the importance of features based on their contribution to distinguishing between different classes, the Relief Algorithm helps in selecting the most informative features for model training. The above Equation 1 shows the Relief-F for attribute A with respect to the target variable T as:

$$W(A) = w[A] - \sum_{i=1}^m \left( diff(A, x_i, H_i) - \frac{1}{k-1} + \sum_{j=1, j \neq class(x_i)}^k diff(A, x_i, M_{ij}) \right)$$

## 1.5 Machine Learning Models

Machine learning algorithms can figure out relationships and patterns from data to make predictions or judgments, without the need for additional programming. This study focused on the identification of fake Instagram profiles using a range of machine learning models, which utilized to accurately classify and detect these profiles.

### 1.5.1 Shallow Models

Shallow models are characterized by their simplicity and ease of interpretation. In this study, we have taken the following shallow models for classification:

#### 1.5.1.1 Support Vector Machine (SVM)

For classification and regression problems, SVM is an effective method that seeks to identify the best hyperplane that divides classes with the largest margin among the nearest points of each class. The Equation 7 for the hyperplane in the context of a SVM expressed as:

$$w \cdot F + b = 0$$

$$\min_{w,b} \frac{1}{2} ||w||^2 \quad \text{subject to } y_i(w \cdot F_i + b) \geq 1, \forall_i$$

#### 1.5.1.2 K-Nearest Neighbors (KNN)

A non-parametric approach taken for problem involving regression as well as classification tasks. It classifies new data points by comparing them to existing data points in the dataset based on their similarity. As shown in Equation 9 that KNN does not have a model in the traditional sense but rather classifies new instances based on a majority vote of the nearest neighbors.

$$\text{Class}(x) =$$

$$\text{mode}\{\text{Class}(x_1), \text{Class}(x_2), \dots, \text{Class}(x_n)\}$$

#### 1.5.1.3 Decision Trees

Tree-like model with internal nodes containing features, branches expressing decision rules depending on them, and leaf nodes indicating outcomes or class labels. Decision trees are ideal for classification and regression because they are simple and interpretable. They can demonstrate numerical and categorical decision-making. As in Equation 10, it splits the dataset into subsets based on an attribute that results in the maximum information gain.

(10)

$$IG(A, D) = Entropy(D)$$

$$- \sum_{j=1}^K \left( \frac{|D_j|}{|D|} \right) Entropy(D_j)$$

(2) Intrinsic Info

(4) From equation

$$r = \frac{\sum(A_i - \bar{A})}{\sqrt{\sum(A_i - \bar{A})^2}}$$

(6) Where  $w[A]$

## 1.5.1.4 Naive Bayes

Probabilistic Bayes' theorem classifier with immense feature independence. Its efficacy and scalability make it popular for text classification, spam filtering, and sentiment analysis. Using input features calculate the class probabilities and predict the most expected class. The Naive Bayes classifier utilizes Bayes' theorem under the assumption of independence between each pair of features, given the class label. The Equation 11 shows the formula for predicting the class  $c$  of a new feature  $F$ .

$$P(C|F) = \frac{P(F|C)P(C)}{P(F)}$$

$$P(c|F) \propto P(c) \prod_{i=1}^{|A|} P(F_i | c)$$

## 1.5.2 Ensemble Models

Ensemble models enhance predictive performance by combining different fundamental models. In the present study, following ensemble models are applied such as:

### 1.5.2.1 Ada Boost

An ensemble learner is a machine-learning algorithm that combines numerous weak classifiers to generate a strong classifier. This can be achieving this by sequentially modifying the weights of incorrectly categorized occurrences in the training dataset. This allows successive weak learners to focus more on the challenging data points that are difficult to classify. The Equation 13 shows its general expression for prediction.

$$w_{i,t+1} = w_{i,t} \cdot \exp(\alpha_t \cdot I(y_i \neq h_t(F_i)))$$

Calculating  $\alpha_t$

$$\alpha_t = \frac{1}{2} \ln \left( \frac{1 - \epsilon_t}{\epsilon_t} \right)$$

### 1.5.2.2 Random Forest

An adaptable approach to training decision trees using ensemble learning. Each tree trained on a randomly selected subset of data and features, and the ultimate prediction is obtained by summing the forecasts from all the trees. Random Forest mitigates overfitting and enhances the reliability of predictions by determining the most frequently occurring class predicted by the individual trees and using it as

the final output. The general Equation no 15 shows the prediction in Random Forest is an average from all decision trees:

$$RF(F) =$$

$$\text{mode}\{Tree_1(F), Tree_2(F), \dots, Tree_i(F)\}$$

### 1.5.2.3 XG Boost

A popular machine learning classification and regression library with optimized gradient boosting. XGBoost successively creates an ensemble of weak learners using gradient boosting to optimize a differentiable loss function and reduce prediction errors. The Equation no 16 shows the objective function at step  $t$ .

$$Obj^{(t)} = \sum_{i=1}^N [loss(y_i, \hat{y}_i^{(t-1)} + f_t(F_i))] +$$

$$Reg(f_t)$$

complexity.

## 1.5.3 Deep Models

By utilizing several layers of neural networks, deep learning models are defines by their capacity to learn intricate patterns from data. This research made use of the following deep learning models:

### 1.5.3.1 Convolutional Neural Network (CNN)

Deep learning model for picture recognition and data presentation that handles integral and categorical data. Filters help CNN convolutional layers retrieve features from input structures and patterns. Cnns use activation functions, pooling layers, and convolutional processes to learn data hierarchies. This makes them suitable for visual tasks including object recognition, image classification, and segmentation, as well as datasets with spatial or structural relationships that affect classification results.

### 1.5.3.2 Recurrent Neural Network (RNN)

The form of neural network framework explicitly created to operate the continuous data by holding the context of the prior input data. Rnns can store information and mould the future predictions by the interlinked network that forms loop inside the networks. There is problem of eliminating gradients when holding the long or

lengthy patterns due to capturing the short-term dependencies of conventional RNN.

### 1.5.3.3 Long Short-Term Memory (LSTM)

LSTM network having continuous data mainly operate at the long-term dependencies. To store and change information over period LSTM have storage cell and gating techniques and capable of handling vanishing gradient problem. Audio recognition, NLP and time series analysis are some of the applications of LSTM.

## 1.6 Experimental Setup

In order to differentiate between the real and fake accounts certain parameters needs to be consider that are cruial for evaluating how well the model performs.This study experimental setup includes the preprocessing dataset, feature

selection, application of machine learning models, and performance evaluation measures.

### 1.6.1 Dataset

The Fake Authenic Instgram User dataset was collected from Kaggle source, a platform for machine learning specialist and data scientists to connect and share their work. The dataset was published with the Attribution 4.0 International License (CC BY 4.0 DEED). The dataset was public on February 11, 2021 by author Kristo Radion Purba. In order to distinguish between real and fake Instagram users the scholars to apply machine-learning classifiers on user profile dataset. The dataset contains the 17 different features having total 65332 users data with two output class lables like fake and real as illustrated in table 1.

Table 1 Dataset Statistics

|                |       |
|----------------|-------|
| Total Accounts | 65332 |
| Real (r)       | 32463 |
| Fake (f)       | 32869 |

### 1.6.2 Performance Metrics

Refers to the criteria used to access the model efficiency, productivity and how well a model is performing the specific task it was designed for. Different metrics are avaiable to evaluate a model's performance. Knowing how these metrics are calculated can help in selecting the most suitable model based on the quantitative measures of performance within those metrics.

- **Accuracy**

Accuracy is a metric that highlights the proportion of the outcomes that were correctly retrieve. It evaluates the performance of information extraction.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

- **Precision**

It evaluates how well the models positive predictions are. Precision focus on the correctness of the positive considering how well the model avoids false positives as shown in equation 18.

$$Precision = \frac{TP}{TP+FP}$$

Recall measures how well the model identifies the positive instances. High recall means model is able to predict most of the positive instances as compared to negative ones. It is important when cost of missing positive is high. The equation 19 shows the calculation for Recall.

$$Recall = \frac{TP}{TP+FN}$$

Combining both precision and recall resulting in measuring F1-Score. Its score ranges from 0 to 1, with 1 being the best. It can be calculated using equation 20.

$$F1\ Score = 2 * \left( \frac{Precision*Recall}{Precision+Recall} \right)$$

The area under the ROC curve (AUC) evaluates the overall performance of the model. Especially useful for evaluating the performance of binary classification models.

$$TPR = \frac{TP}{TP+FN}$$

## 2 Results And Discussion

The classification models were evaluated using various performance metrics, including accuracy, precision, recall, and F1-score. The results of different machine learning techniques are presented in the following subsections.

### 2.1 Parameters Settings

The classification is done using the spyder IDE (5.4.3) and python libraries such as pandas, numpy, matplotlib, tensorflow, sklearn, seaborn, and xgboost are applied. The dataset is splitted 70% for training and 30% for testing. Python-based machine learning techniques are applied under different configurations. The SVM was configured to use the default kernel and probability calculations. The Naive Bayes model was used with default settings. The Decision Tree classification used a Gini score and a random state of 0. KNN was adjusted to use 30 neighbors. The Random Forest classifier employed 100 estimators and a random state of 15. XG Boost and adaboost were set normally. For more complex modeling, a CNN with 64 and 128 filters, pool size 2, a thick layer of 256 units,

dropout at 0.5, and rmsprop optimizer was utilized. The RNN and LSTM models contained 64 and 128 units, 256 units in a thick layer, and 0.5 dropout rates. Adam enhanced their performance. These parameters optimized each model for data and study outcomes.

### 2.2 Data Visualization

Let us discuss the data visualization and experimental results of Fake Authentic Instagram User dataset. Performing data visualization on using 17 dataset features, the heatmap as illustrated in Figure 3 provides a visual representation of the correlation between different features in your dataset. The coefficients can take on values between -1 and 1, with 1 indicating a highly positive correlation, -1 a highly negative correlation, and 0 indicating no association at all. When two variables go up or down together, it's because of a strong positive correlation. Any significant negative correlations would imply that an increase in one feature corresponds to a decrease in another.

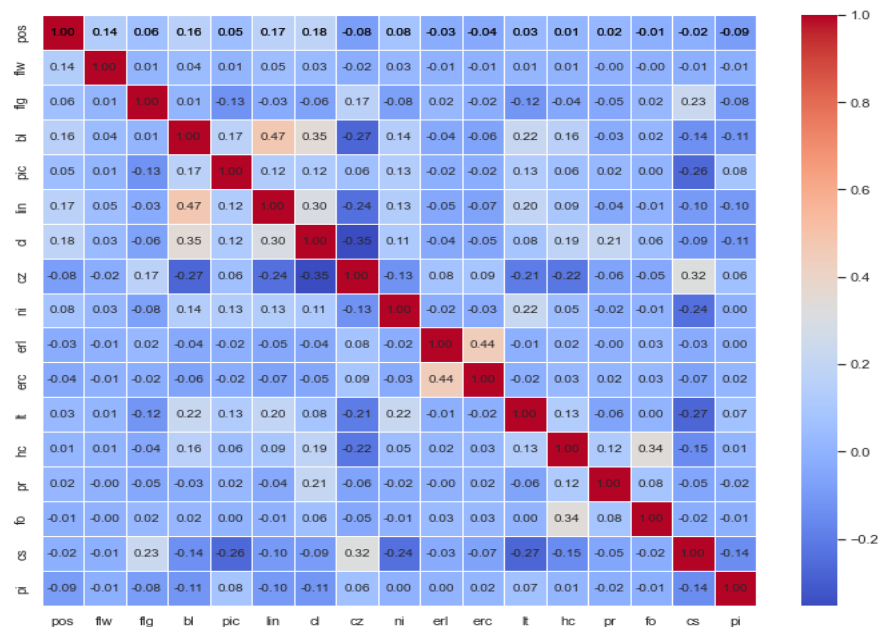


Figure 3: Heatmap Diagram Showing Correlation between Dataset Features

By examining the correlations between different user metrics, interesting patterns and behaviors might be revealed. A moderate positive correlation of 0.14 between number of posts (pos) and followers (flw) implies that active users who publish often follow more users. Number of Followers (flg) positively correlate with engagement rate likes (erl) by 0.17, suggesting that persons with more followers get more likes on their posts, which may reflect popularity or engagement.

Picture availability (pi) appears to be positively connected with biography length (bl). This may suggest that people who write extensive biographies are more likely to have profile pictures. The picture availability (pic) has a 0.03 connection with zero-length captions (cz), indicating that a profile image does not necessarily correspond with textual content in posts. However, zero-length captions are negatively correlated with average caption length (cl) by -0.35. It shows that longer captions are less likely to be empty, indicating a user's intent to engage their audience. Non-image percentage (ni) and engagement rate likes have a -0.13 visible negative correlation. It suggests that image-based postings may get more likes than text-only posts,

confirming social media's visual character. However, there is no significant correlation with comment engagement rate (erc), suggesting that while images are more likely to be liked, they may not encourage more commentary.

Post interval (pi) is negatively correlated (-0.09) with number of post (pos), showing that frequent posters post more often. This trend might either overwhelm followers with content, causing engagement fatigue, or maintain a regular stream of communication, which could boost visibility and interaction.

## 2.3 Top 10 Ranked Features

In this research top 10 features are extracted from the dataset. The table 2 presents a comparison of feature rankings using four different methods such as Correlation, Relief-F, Information Gain (IG), and Gain Ratio (GR). This analysis assists in determining which features have the highest predictive performance for the desired outcomes. Notably, the feature 'flg' consistently ranks highest across all methods, underscoring its robust predictive performance and importance across various statistical measures

Table 2 Feature Ranking By IG, GR, Relief-F and Correlation

| Correlation |                 |          | Relief-F        |          | Information Gain |          | Gain Ratio      |          |
|-------------|-----------------|----------|-----------------|----------|------------------|----------|-----------------|----------|
| Sr. No      | Ranked Features | Values   | Ranked Features | Values   | Ranked Features  | Values   | Ranked Features | Values   |
| 1.          | Flg             | 0.396000 | Cs              | 0.084498 | Erc              | 0.330871 | Lin             | 0.119540 |
| 2.          | Lin             | 0.368177 | Flg             | 0.036476 | Flg              | 0.158288 | Pic             | 0.083849 |
| 3.          | Lt              | 0.320200 | Lt              | 0.034119 | Lin              | 0.103985 | Pi              | 0.062931 |
| 4.          | Cs              | 0.306951 | Cz              | 0.032488 | Erl              | 0.077611 | Cs              | 0.054835 |
| 5.          | Bl              | 0.254126 | Ni              | 0.029798 | Flw              | 0.072563 | Lt              | 0.029137 |
| 6.          | Cz              | 0.204169 | Erc             | 0.026191 | Pi               | 0.042088 | Pr              | 0.028572 |
| 7.          | Pic             | 0.169713 | Hc              | 0.019807 | Lt               | 0.031771 | Erl             | 0.021476 |
| 8.          | Ni              | 0.161461 | Bl              | 0.017945 | Pos              | 0.031518 | Erc             | 0.018780 |
| 9.          | Pi              | 0.155808 | Pr              | 0.016031 | Cl               | 0.031151 | Flg             | 0.017840 |
| 10.         | Pr              | 0.104142 | Cl              | 0.014404 | Bl               | 0.028780 | Bl              | 0.014057 |

Similarly, 'pi' and 'bl' appear prominently across the methods, indicating their utility in the model. These features maintain high rankings, especially

in Information Gain and Gain Ratio, pointing to their capability in effectively splitting the data and reducing entropy. Their performance

suggests that they can significantly impact model decisions and should be included in the feature set for model training.

The Relief-F technique assesses how well feature distinguish its neighbour, highlighting 'pos' and 'flw', which may be important for algorithms sensitive to neighbour structure such as KNN. However, features such as "pr," "lt," and "pic," which regularly score low across these evaluation criteria, doesn't effect much on the target variable. Their lower rankings suggest that

including these features might not contribute significantly to model performance.

## 2.4 Feature Wise Classification Results

On top ranked attributes of dataset based on feature ranking techniques above mentioned in Table 2, machine learning classifiers SVM, Navie bayes, DT, KNN, RF, XG Boost, adaboost, CNN, RNN and LSTM are applied on them. Sklearn model\_selection python libaray is used to split data for testing and training. Accuracy, Precsion, Recall, F-measure and ROC are used a performance evaluation measure.

Table 3 Classification using Correlation and Relief-F (results in %)

|                      | ML Algorithms | Correlation |       |       |       | Relief-F |       |       |       |
|----------------------|---------------|-------------|-------|-------|-------|----------|-------|-------|-------|
|                      |               | Acc         | Pre   | Rec   | F1    | Acc      | Pre   | Rec   | F1    |
| Shallow Models       | SVM           | 81.71       | 78.75 | 86.77 | 82.57 | 83.60    | 80.94 | 87.82 | 84.24 |
|                      | Naive Bayes   | 74.24       | 68.53 | 89.50 | 77.63 | 68.47    | 61.99 | 95.24 | 75.10 |
|                      | DT            | 75.25       | 75.22 | 75.19 | 75.21 | 81.92    | 82.69 | 80.68 | 81.67 |
|                      | KNN           | 81.22       | 78.87 | 85.21 | 81.92 | 82.50    | 79.43 | 87.65 | 83.34 |
| Ensemble Models      | Random Forest | 81.91       | 80.20 | 84.66 | 82.37 | 87.68    | 83.64 | 93.64 | 88.36 |
|                      | XG Boost      | 82.17       | 80.20 | 85.34 | 82.69 | 87.93    | 83.95 | 93.74 | 88.58 |
|                      | Ada Boost     | 80.79       | 79.88 | 82.23 | 81.04 | 85.54    | 83.43 | 88.64 | 85.96 |
| Deep Learning Models | CNN           | 81.46       | 79.23 | 85.21 | 82.11 | 86.36    | 81.54 | 93.93 | 87.30 |
|                      | RNN           | 80.89       | 78.16 | 85.65 | 81.74 | 86.27    | 81.32 | 94.12 | 87.26 |
|                      | LSTM          | 79.23       | 74.03 | 89.96 | 81.22 | 80.06    | 75.73 | 88.40 | 81.58 |

The outperforming feature in terms of Correlation and Relief-F is xgboost with an accuracy of 82.17% and 87.93% dominated all machine learning classifiers as shown in Table 3. Indicating its robustness and effectiveness in handling the given dataset and feature selection

techniques. The Table 4 shows that incase of Information Gain Random forest dominates with accuracy of 88.82% and incase of Gain Ratio xgboost achieves an overall accuracy of approximately 82.17%.

Table 4 Classification using Information Gain and Gain Ratio

|                | ML Algorithm  | Information Gain |       |       |       | Gain Ratio |       |       |       |
|----------------|---------------|------------------|-------|-------|-------|------------|-------|-------|-------|
|                |               | Acc              | Pre   | Rec   | F1    | Acc        | Pre   | Rec   | F1    |
| Shallow Models | SVM           | 84.95            | 81.99 | 89.52 | 85.59 | 74.24      | 68.53 | 89.50 | 77.63 |
|                | Naive Bayes   | 71.89            | 65.79 | 91.04 | 76.38 | 75.25      | 75.22 | 75.19 | 75.21 |
|                | DT            | 83.88            | 84.37 | 83.09 | 83.73 | 81.71      | 78.75 | 86.77 | 82.57 |
|                | KNN           | 83.60            | 80.68 | 88.29 | 84.31 | 81.22      | 78.87 | 85.21 | 81.92 |
| Ensemble       | Random Forest | 88.82            | 84.85 | 94.48 | 89.40 | 81.79      | 79.92 | 84.83 | 82.30 |

|                      |           |       |       |       |       |       |       |       |       |
|----------------------|-----------|-------|-------|-------|-------|-------|-------|-------|-------|
| Models               | XG Boost  | 88.80 | 85.30 | 93.72 | 89.32 | 82.17 | 80.21 | 85.34 | 82.70 |
|                      | Ada Boost | 85.47 | 84.47 | 86.87 | 85.65 | 80.79 | 79.88 | 82.23 | 81.04 |
| Deep Learning Models | CNN       | 86.65 | 81.76 | 94.29 | 87.58 | 81.56 | 79.19 | 85.55 | 82.24 |
|                      | RNN       | 86.45 | 80.94 | 95.29 | 87.53 | 81.47 | 79.70 | 84.36 | 81.97 |
|                      | LSTM      | 83.38 | 80.11 | 88.75 | 84.21 | 80.11 | 78.38 | 81.28 | 80.32 |

## 2.5 Categorical Classification Results

To effectively handle and analyze large complex dataset organizing seventeen features into five distinct categories related to user profile information, content characteristics, interaction metrics, and content similarity measures. This classification enables focused analysis, where each set of Features has similar characteristics related to user behavior and content engagement. This simplifies the modeling process for identifying fake profiles. Applying machine learning algorithms to each unique category to better predict fake profiles using specific feature sets and to better improve model accuracy and efficiency.

### 2.5.1 User Profile Information

It provides insights into the static attributes of users, such as their social connectivity and personal description. The table 5 shows that XG Boost showed the best performance in this category with an accuracy of 84.64%. This suggests that XG Boost is particularly effective at using these profile-based features to classify user information, possibly due to its robustness against overfitting and ability to handle diverse feature types, from numerical counts (like posts, followers, and following) to binary features (such as picture and link availability).

Table 5: User Profile Information Classification

| Category                 | ML Algo     | Accuracy     | Precision    | Recall       | F1-Score     |
|--------------------------|-------------|--------------|--------------|--------------|--------------|
| User Profile Information | SVM         | 77.21        | 74.83        | 73.65        | 75.39        |
|                          | Naive Bayes | 20.26        | 58.21        | 83.63        | 32.61        |
|                          | DT          | 71.82        | 72.74        | 73.11        | 72.46        |
|                          | KNN         | 80.52        | 76.83        | 74.94        | 77.63        |
|                          | RF          | 81.96        | 79.06        | 77.42        | 79.62        |
|                          | XG Boost    | <b>84.64</b> | <b>80.14</b> | <b>77.61</b> | <b>80.97</b> |
|                          | Ada Boost   | 81.24        | 79.14        | 77.92        | 79.55        |
|                          | CNN         | 81.27        | 74.81        | 71.92        | 76.31        |
|                          | RNN         | 88.35        | 74.68        | 69.34        | 77.70        |
|                          | LSTM        | 69.27        | 73.48        | 75.57        | 72.28        |

### 2.5.2 Interaction Metrics

Analyzes the level of user involvement by looking at the number of likes and comments, which can serve as indicators of the content's popularity and influence. The table 6 illustrates that SVM performed best here approximately 72.69%. The efficacy of SVM may arise from its capacity to

detect similarities in interaction patterns among users, which is crucial in datasets where interaction metrics differ significantly across various user accounts.

Table 6: Interaction Metrics Classification

| Category            | ML Algo       | Accuracy | Precision | Recall | F1-Score |
|---------------------|---------------|----------|-----------|--------|----------|
| Interaction Metrics | SVM           | 72.69    | 69.34     | 68.07  | 70.30    |
|                     | Naive Bayes   | 96.90    | 51.05     | 50.51  | 66.41    |
|                     | Decision Tree | 59.39    | 62.45     | 63.18  | 61.23    |
|                     | KNN           | 68.21    | 71.54     | 73.01  | 70.53    |
|                     | Random Forest | 65.34    | 64.81     | 64.59  | 64.96    |
|                     | Xgboost       | 67.61    | 71.40     | 73.08  | 70.24    |
|                     | Adaboost      | 66.11    | 70.23     | 71.98  | 68.92    |
|                     | CNN           | 70.32    | 70.21     | 70.66  | 70.49    |
|                     | RNN           | 69.98    | 70.79     | 71.07  | 70.52    |
|                     | LSTM          | 68.22    | 70.03     | 70.72  | 69.45    |

**Tagging And Hashtags**

Determining the contextual relevance and categorizing the user content involves analysis of the metadata. The table 7 shows that the best performer in terms of tagging and hashtag was the CNN deep learning algorithm having the best

accuracy rate of 71.21%. The textural data like the has location and hastags having which may differ in both context and formatting, the convolution nature of CNN is well suited in analyzing and recognizing their pattern.

Table 7: Tagging and hashtags Classification

|                      | ML Algo       | Accuracy | Precision | Recall | F1-Score |
|----------------------|---------------|----------|-----------|--------|----------|
| Tagging and Hashtags | SVM           | 66.10    | 67.09     | 67.37  | 66.73    |
|                      | Naive Bayes   | 42.66    | 63.37     | 72.68  | 53.76    |
|                      | Decision Tree | 59.54    | 65.82     | 68.02  | 63.49    |
|                      | KNN           | 60.51    | 66.25     | 68.28  | 64.16    |
|                      | Random Forest | 61.10    | 66.25     | 68.04  | 64.38    |
|                      | Xgboost       | 63.39    | 66.91     | 68.12  | 65.67    |
|                      | Adaboost      | 60.24    | 67.27     | 70.01  | 64.76    |
|                      | CNN           | 71.21    | 67.71     | 67.03  | 69.06    |
|                      | RNN           | 68.19    | 67.35     | 67.00  | 67.59    |
|                      | LSTM          | 57.90    | 66.80     | 70.36  | 63.52    |

**Content Characteristics**

Examines the nature of the content posted by users, focusing on textual features and media types, which can indicate content quality and user engagement. The table 8 illustrate that CNN was the top performer in this category with an accuracy of 67.75%. Cnns are typically adapt at

processing data with spatial hierarchies and may be effectively capturing patterns in text length and content format, making them suitable for analyzing content characteristics.

Table 8: Content Characteristics Classification

| Category                | ML Algo     | Accuracy | Precision | Recall | F1-Score |
|-------------------------|-------------|----------|-----------|--------|----------|
| Content Characteristics | SVM         | 67.23    | 67.49     | 67.52  | 67.37    |
|                         | Naive Bayes | 77.64    | 58.59     | 56.17  | 65.18    |
|                         | DT          | 55.45    | 61.34     | 62.77  | 58.88    |

|               |       |       |       |       |
|---------------|-------|-------|-------|-------|
| KNN           | 67.76 | 67.10 | 66.81 | 67.28 |
| Random Forest | 61.61 | 63.20 | 63.55 | 62.57 |
| XG Boost      | 69.98 | 67.34 | 66.40 | 68.15 |
| Ada Boost     | 70.78 | 67.63 | 66.52 | 68.58 |
| CNN           | 67.75 | 67.77 | 68.30 | 68.02 |
| RNN           | 66.57 | 67.98 | 68.43 | 67.49 |
| LSTM          | 65.01 | 67.85 | 68.85 | 66.88 |

## Content Similarity Measures

For determining whether the profile is fake or not can also be determined by the content similarity due to the presence of repeated content. The table 9 shows that the best performer for content similarity measures is the SVM machine learning

algorithm, having the best accuracy of 83.84%. Differentiating between the complex data patterns that involve the content similarity can be resolved using SVM due to its nature of handling the large sized data.

**Table 9: Content Similarity Measures Classification**

| Category                    | ML Algo       | Accuracy | Precision | Recall | F1-Score |
|-----------------------------|---------------|----------|-----------|--------|----------|
| Content Similarity Measures | SVM           | 83.84    | 66.04     | 61.78  | 71.14    |
|                             | Naive Bayes   | 91.88    | 64.16     | 59.07  | 71.91    |
|                             | Decision Tree | 58.51    | 61.54     | 62.21  | 60.30    |
|                             | KNN           | 65.95    | 66.98     | 67.27  | 66.60    |
|                             | Random Forest | 63.98    | 64.15     | 64.13  | 64.06    |
|                             | Xgboost       | 70.77    | 67.23     | 66.02  | 68.31    |
|                             | Adaboost      | 76.85    | 66.84     | 63.98  | 69.82    |
|                             | CNN           | 83.40    | 65.81     | 61.65  | 70.89    |
|                             | RNN           | 82.67    | 65.71     | 61.68  | 70.65    |
|                             | LSTM          | 76.34    | 66.97     | 64.23  | 69.76    |

These results help us in determining the fake profiles. Fake accounts have unusual followers and user bios which are detected in user profile category. Interaction metrics and tagging hashtag can reveal unauthentic connection trends, like improper use of hashtags and tagging users or high or low interaction metrics. Recognizing the repetitive advertising messages or similar posts across several accounts have been identified using content characteristics and similarity measures. These methods are often used by fake profiles to gain attention.

## Categorical Combinations Classification Results

The Instagram users' behaviours can extensively be determined by integrating the multiple dataset features categories as explained in above section 4.5. They identify the irregularities and

inconsistencies between different patterns like interactions, social connections, and user contents of the fake Instagram profiles. To reduce the risk of overfitting for various dataset features combining different categories of user activities and their profile features. This enhanced the reliability and accuracy of the different classification algorithms making the process more secure against the various different techniques used by fake profiles.

## User Profile Information & Interaction Metrics

The highest performing machine learning algorithm in the following combination was Random Forest as shown in table 10, having an accuracy rate of 89.24%. High precision and recall indicate the model's effectiveness in accurately identifying fake profiles without many

false positives, which is crucial for maintaining user trust.

Table 10: User Profile Info and Interaction Metrics Classification

| Category Combination                           | ML Algo       | Accuracy | Precision | Recall | F1-Score |
|------------------------------------------------|---------------|----------|-----------|--------|----------|
| User Profile Information & Interaction Metrics | SVM           | 84.66    | 80.92     | 90.64  | 85.50    |
|                                                | Naive Bayes   | 67.67    | 62.91     | 85.81  | 72.60    |
|                                                | Decision Tree | 85.19    | 85.62     | 84.53  | 85.07    |
|                                                | KNN           | 85.65    | 81.40     | 92.37  | 86.54    |
|                                                | Random Forest | 89.24    | 85.35     | 94.70  | 89.78    |
|                                                | XG Boost      | 89.18    | 85.52     | 94.30  | 89.69    |
|                                                | Ada Boost     | 86.01    | 85.71     | 86.39  | 86.05    |
|                                                | CNN           | 86.25    | 80.38     | 95.86  | 87.44    |
|                                                | RNN           | 85.90    | 79.23     | 97.25  | 87.32    |
|                                                | LSTM          | 85.64    | 79.15     | 96.70  | 87.05    |

#### Content Characteristics & Content Similarity Measures

The table 11 illustrates that the XG Boost model performs slightly better than others, achieving an accuracy of 71.88% and an F1-Score of 71.99%.

The ability to analyze content depth and similarity aids in detecting profiles that might use repetitive content, which is common in fake profiles.

Table 11: Content Characteristic Classification and Similarity Measures

| Category Combination                          | ML Algo       | Accuracy | Precision | Recall | F1-Score |
|-----------------------------------------------|---------------|----------|-----------|--------|----------|
| Content Characteristics & Similarity Measures | SVM           | 71.25    | 69.17     | 76.53  | 72.66    |
|                                               | Naive Bayes   | 65.88    | 60.93     | 88.25  | 72.09    |
|                                               | Decision Tree | 64.24    | 64.79     | 62.11  | 63.42    |
|                                               | KNN           | 70.72    | 69.34     | 74.12  | 71.65    |
|                                               | Random Forest | 70.35    | 69.97     | 71.17  | 70.56    |
|                                               | XG Boost      | 71.88    | 71.60     | 72.39  | 71.99    |
|                                               | Ada Boost     | 70.87    | 69.46     | 74.34  | 71.81    |
|                                               | CNN           | 71.31    | 70.84     | 72.29  | 71.56    |
|                                               | RNN           | 71.55    | 70.64     | 73.59  | 72.09    |
|                                               | LSTM          | 71.31    | 69.77     | 75.03  | 72.31    |

#### User Profile Information & Tagging And Hashtags

In this category XG Boost leads again with an accuracy of 81.55% and F1-Score of 82.28%

shown in table 12. This suggests that combining user profile details with tagging habits provides strong indicators for distinguishing authentic from fake user engagements.

Table 12: User Profile Info, Tagging and hashtags Classification

| Category Combination                               | ML Algo       | Accuracy | Precision | Recall | F1-Score |
|----------------------------------------------------|---------------|----------|-----------|--------|----------|
| User Profile Information<br>& Tagging and Hashtags | SVM           | 77.84    | 77.77     | 77.86  | 77.82    |
|                                                    | Naive Bayes   | 59.17    | 84.17     | 22.44  | 35.44    |
|                                                    | Decision Tree | 75.02    | 75.05     | 74.84  | 74.94    |
|                                                    | KNN           | 78.78    | 77.11     | 81.76  | 79.37    |
|                                                    | Random Forest | 81.08    | 78.96     | 84.67  | 81.72    |
|                                                    | XG Boost      | 81.55    | 79.04     | 85.79  | 82.28    |
|                                                    | Ada Boost     | 80.14    | 79.46     | 81.19  | 80.32    |
|                                                    | CNN           | 77.93    | 78.25     | 77.27  | 77.76    |
|                                                    | RNN           | 77.87    | 78.52     | 76.64  | 77.57    |
|                                                    | LSTM          | 76.92    | 77.82     | 75.19  | 76.48    |

### Interaction Metrics & Content Similarity Measures

Interaction metrics provide insight into how users interact with the content. Abnormally high or low interaction rates can signal automated or non-genuine activities often seen in bot-driven or commercially-driven fake profiles. High similarity

scores might indicate repetitive, or copied content which is common in fake profiles. So table 13 shows that XG Boost show comparable high performances with accuracies around 70.69% for the analysis of fake or real profile data.

Table 13: Interaction Metrics and Content Similarity Measures Classification

| Category Combination                                  | ML Algo       | Accuracy | Precision | Recall | F1-Score |
|-------------------------------------------------------|---------------|----------|-----------|--------|----------|
| Interaction Metrics<br>Content Similarity<br>Measures | SVM           | 70.62    | 69.99     | 72.07  | 71.01    |
|                                                       | Naive Bayes   | 64.74    | 59.58     | 91.35  | 72.12    |
|                                                       | Decision Tree | 62.93    | 63.50     | 60.57  | 62.00    |
|                                                       | KNN           | 69.18    | 68.08     | 72.06  | 70.01    |
|                                                       | Random Forest | 68.48    | 67.97     | 69.72  | 68.83    |
|                                                       | XG Boost      | 70.69    | 70.81     | 70.24  | 70.52    |
|                                                       | Ada Boost     | 70.17    | 69.51     | 71.70  | 70.59    |
|                                                       | CNN           | 70.55    | 69.72     | 72.51  | 71.09    |
|                                                       | RNN           | 70.34    | 69.42     | 72.55  | 70.95    |
|                                                       | LSTM          | 69.84    | 71.10     | 66.70  | 68.83    |

### Content Characteristics, Tagging & Hashtags, And Interaction Metrics

Detailed captions and a variety of media(image, video) indicate authenticity whereas fake profile users may share less diverse images and adopt short, repetitive captions. Excessive or

inappropriate hashtag especially for promotion or followers, may indicate spam or fake profile. The table 14 shows that Random Forest having an accuracy of 79.94% performs best for the analysis of fake profile using the machine learning algorithms.

Table 14: Interaction Metrics and Content Characteristics and Tagging Hashtag Classification

| Category Combination                                | ML Algo     | Accuracy | Precision | Recall | F1-Score |
|-----------------------------------------------------|-------------|----------|-----------|--------|----------|
| Content Characteristics,<br>Tagging & Hashtags, and | SVM         | 73.27    | 73.31     | 73.06  | 73.18    |
|                                                     | Naive Bayes | 55.54    | 53.24     | 89.83  | 66.86    |

|                     |               |       |       |       |       |
|---------------------|---------------|-------|-------|-------|-------|
| Interaction Metrics | Decision Tree | 70.52 | 70.83 | 69.60 | 70.21 |
|                     | KNN           | 73.04 | 72.57 | 73.97 | 73.26 |
|                     | Random Forest | 76.94 | 76.40 | 77.84 | 77.12 |
|                     | XG Boost      | 77.01 | 76.16 | 78.55 | 77.33 |
|                     | Ada Boost     | 74.70 | 75.48 | 73.04 | 74.24 |
|                     | CNN           | 74.30 | 72.04 | 79.30 | 75.50 |
|                     | RNN           | 74.46 | 73.33 | 76.76 | 75.01 |
|                     | LSTM          | 72.45 | 73.01 | 71.11 | 72.04 |

According to results from numerous model trials, machine learning models, such as Random Forest and XG Boost, exhibit robust performance and high accuracy in identifying false profiles when analyzing integrated features from distinct dataset categories. The effectiveness of these models in various combinations suggests that a multi-faceted approach to data analysis can significantly enhance the detection of fake profiles. As a result, these models strengthen the safety and reliability of social media platforms by spotting anomalies that could go undetected when

focusing on a particular feature category.

#### Classification Results On All Features

Accuracy, Precision, Recall, and F1-Score are important measures to assess the performance of different machine learning algorithms in two-class and four-class datasets classification results. In the classification of all features, table 15 shows that Random Forest performs well having an accuracy of 90.10% is particularly notable. XG Boost also competes against the RF with an accuracy of 89.96%.

Table 15: All Dataset Features Classification Results

| ML Algorithms | Accuracy | Precision | Recall | F1-Score |
|---------------|----------|-----------|--------|----------|
| SVM           | 85.27    | 82.57     | 89.36  | 85.83    |
| Naive Bayes   | 73.45    | 79.94     | 62.71  | 70.28    |
| Decision Tree | 86.56    | 86.26     | 84.55  | 85.39    |
| KNN           | 83.06    | 80.11     | 87.89  | 83.82    |
| Random Forest | 95.10    | 92.15     | 94.51  | 90.59    |
| XG Boost      | 89.96    | 86.78     | 94.26  | 90.36    |
| Ada Boost     | 86.85    | 86.44     | 87.37  | 86.90    |
| CNN           | 86.72    | 82.43     | 93.04  | 87.41    |
| RNN           | 86.38    | 82.64     | 91.80  | 86.98    |
| LSTM          | 82.62    | 78.82     | 88.77  | 83.50    |

#### Receiver Operating Characteristic (Roc) Curve

Receiver Operating Characteristic curves examine the sensitivity and precision of various machine-learning algorithms at different threshold values. By contrasting the false positive and true positive rates, evaluating how well they perform. The higher AUC illustrates the algorithm's ability to distinguish between the classes, indicating the model's superior performance. False positive and

True positive are shown on x and y axes respectively.

#### Roc Curves For Different Classifiers

The Figure 4 (a) illustrates the ROC curves for traditional methods such as SVM, Naive Bayes, DT, and KNN. The curves demonstrate that SVM achieves the highest AUC with a value of 0.92, and then follows the KNN with 0.88, DT with 0.85, and Naive Bayes with the lowest value

of 0.83. This performance range shows how different these models can be in terms of efficacy based on the kind of data they handle and their underlying mechanics.

ROC curves for CNN, RNN, and LSTM deep learning models are shown in The Figure 4 (b). AUC values for CNN and RNN are about 0.93, while LSTM values are a little lower at 0.91, all three models show a strong ability to discriminate across classes. The steep curves in all models

towards the upper left corner show good sensitivity and specificity balance.

The The Figure 4 (c) analyze the performance of ensemble models RF, XG Boost, and Ada Boost by illustrating the ROC curves. Aucs of 0.96 for RF and XG Boost and 0.94 for Ada Boost show outstanding predictive performance. These models work well by combining several weak learners to create a powerful predictive model, improving performance.

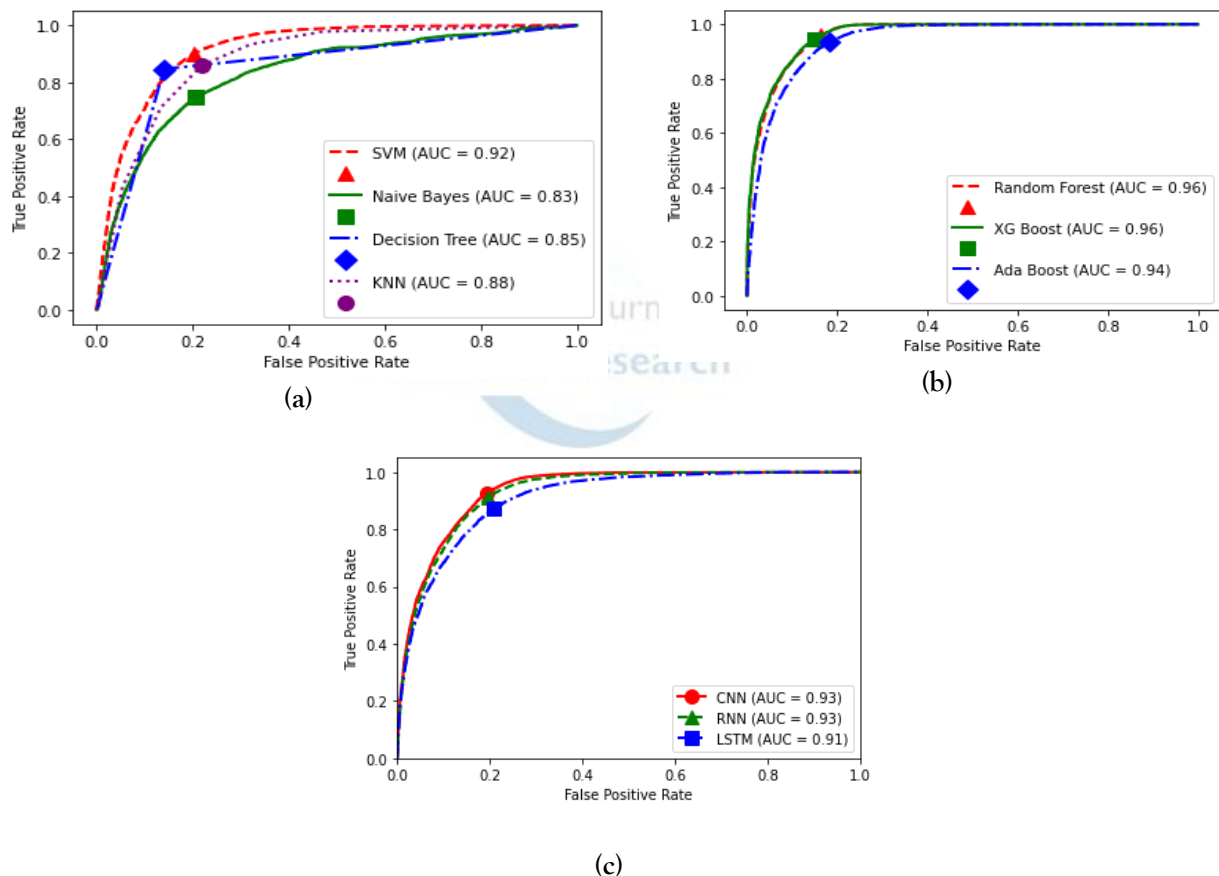


Figure 4: ROC Curves for applied algorithms (a) Shallow ML (b) Ensemble Models (c) Deep Models

### Optimal Results With Existing Studies

Different machine learning classifiers applied to the dataset features individual, category wise, combination of categories, and top 10 features selected using feature selection techniques. Looking at the overall results of the dataset, and getting the top performing models, shallow model SVM shows strong performance with an

accuracy of 84.59%, the ensemble-learning model Random Forest achieves an accuracy of 95.10%, and deep learning model CNN achieves an accuracy of 86.36%. The top predictive feature among these models includes the comments engagement rate ('erc'), link availability ('lin'), and number of followers ('flg') highlighting the importance of user profile information and

interaction. Concluding that the Random Forest the ensemble-learning model outperforms other having an impressive accuracy of 95.10% and the user profile information and interaction metrics of user profile plays an important role in Instagram user profile analytics.

For the classification of Instagram profiles, either as fake or real profile different studies made over the past few years using machine-learning algorithms as shown in table 16. Exploring various existing studies for the methods applied in order to classify profiles. In 2019 SVM, Naive Bayes, and RF algorithms applied achieving an accuracy of 86% and 92% respectively. LSTM and Deep Neural Networks are applies together resulting in accuracy of 86% in 2020. Continuing the work in 2022 again, the combination of conventional SVM and Random Forest

algorithms classify profiles with an accuracy rate of 91%. The recent work in 2023 applied a Novel Integrated Neural Network algorithm having an accuracy of 85.45%.

Our study for the classification of fake instagram profile involves the applying the machine-learning models like shallow, ensemble and deep learning algorithms. Applying these algorithms on dataset features individually, categorical, combination of these categories and on the top10 selected features using features selection techniques. Our experimental results shows that an ensemble-learning model Random Forest outperforms the existing studies and other classification models that have been applied achieves an impressive accuracy of 95.10% indicating a potential improvement in the classification of fake Instagram profiles.

**Table 16: The comparison of all models with proposed model (results in %)**

| Ref      | Year | Algorithm            | Results (Acc) |
|----------|------|----------------------|---------------|
| [12]     | 2019 | Random Forest        | 90            |
| [20]     | 2020 | DNN, LSTM            | 86            |
| [9]      | 2022 | SVM, Naive Bayes.    | 83.5          |
| [19]     | 2023 | Naive Bayes, SVM     | 91.2          |
| [21]     | 2023 | Novel Integrated NN  | 90.1          |
| [15]     | 2024 | Deep Neural Networks | 92.9          |
| [22]     | 2025 | Decision Tree, SVM   | 93.17         |
| Proposed | 2026 | Random Forest        | 95.10         |

### Conclusions And Future Research Directions

This research study conducts a comprehensive analysis of the fake authentic Instagram user dataset to evaluate the classification performance of several classifiers. Multiple categories are considered based on the user's personal information, content characteristics, interaction metrics, tagging/hashtags, and content similarity measures, as well as the combination of these categories. Previous studies typically used user profile data to identify fake profiles. The primary goal of these categories is to improve classifier classification accuracy and effectively identify fake profiles.

In our findings, this study involves analyzing various classifiers, which include shallow machine

learning models like SVM, Naive Bayes, DT, and KNN, as well as ensemble models such as RF, adaboost, and xgboost. Deep learning models like CNN, RNN, and LSTM were also considered. The classifiers were assessed using a dataset of Instagram profiles, which were classified into binary (real or fake) labels. To investigate the effectiveness of these algorithms on dataset features. Features Selection done by selecting the top 10 features from all 17 dataset features that may have the most predictive power are ranked using feature selection techniques like correlation, information gain, gain ratio, and Relief-F. Following that, the classifiers were applied to these selected features in order to

ascertain which set of attributes would result in the maximum level of detection accuracy.

The top-performing model across the feature combinations and classification tasks remained Random Forest. For dataset classification, Random Forest achieves an accuracy of 95.10%. It's ensemble nature, which aggregates predictions from several decision trees, helped manage feature interactions effectively and prevent overfitting. Comment Engagement Rate, Biography length, & Num following and in terms of features category User profile information and Interaction metrics are more predictive dataset features.

These results highlight the importance of feature engineering, appropriate model selection, and data preprocessing for effectively detecting fake profiles on social media. By combining features from user profiles to interaction metrics, models can identify suspicious profiles more accurately. In the future research, divers datasets are explored for social media profile features for the detection of spam bots, and hackers accounts. Using multimodal approaches that combine textual, and behavioral features to improve accuracy and exploring Graph based approaches to detect fake profile anomalies more effectively.

## REFERENCES

- A. Ahaskar, "How fake profiles on social media can take you for a ride," 2019. <https://www.livemint.com/technology/tech-news/how-fake-profiles-on-social-media-can-take-you-for-a-ride-1563466419127.html>
- Stephanie Walker, "Fighting Fake Profiles: Effective Profile Moderation," Chekkee, 2023. <https://chekkee.com/fighting-fake-profiles-effective-profile-moderation/>
- Bytescare, "What is fake profile? Why it is created?," 2023. <https://www.linkedin.com/pulse/what-fake-profile-bytescare-f9ytc/>

- V. Dencheva, "Share of fraudulent Instagram influencers worldwide Jun 28, 2023," 2023. <https://www.statista.com/statistics/1250681/share-of-instagram-influencers-involved-in-fraud-worldwide/>
- R. Shewale, "Instagram Fake Profile Statistics," 2023. <https://www.basichinking.com/bots-instagram-accounts-fake/>
- Santander, "Four steps to identify fake accounts on social media," 2023. <https://www.santander.com/en/stories/four-steps-to-identify-fake-accounts-on-social-media>
- M. Mohammadrezaei, M. E. Shiri, and A. M. Rahmani, "Identifying Fake Accounts on Social Networks Based on Graph Analysis and Classification Algorithms," Secur. Commun. Networks, vol. 2018, 2018, doi: 10.1155/2018/5923156.
- R. R. Rostami and S. Karbasi, "Detecting fake accounts on twitter social network using multi-objective hybrid feature selection approach," Webology, vol. 17, no. 1, 2020, doi: 10.14704/WEB/V17I1/A204.
- 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. 1-12, 2022, doi: 10.1002/cpe.7349.
- J. Gangan, S. KS, N. Jamdar, and S. Bharne, "Detection of Fake Twitter Accounts Using Ensemble Learning Model," 2023 7th Int. Conf. Comput. Commun. Control Autom., pp. 1-6, 2024, doi: 10.1109/iccubea58933.2023.10391985.
- S. Sheikhi, "An efficient method for detection of fake accounts on the instagram platform," Rev. d'Intelligence Artif., vol. 34, no. 4, pp. 429-436, 2020, doi: 10.18280/ria.340407.

- A. Dey, H. Reddy, M. Dey, and N. Sinha, "Detection of Fake Accounts in Instagram Using Machine Learning," *Int. J. Comput. Sci. Inf. Technol.*, vol. 11, no. 5, pp. 83–90, 2019, doi: 10.5121/ijcsit.2019.11507.
- M. VAROL ARISOY and T. TUNÇ ABUBAKAR, "A Comparative Analysis of Ensemble Learning Methods on Social Media Account Detection," *J. Eng. Technol. Appl. Sci.*, vol. 8, no. 2, pp. 87–105, 2023, doi: 10.30931/jetas.1325483.
- D. B. Naik, S. Navya Sri, G. Meenakshi Priya, and P. Satya Priya, "Machine Learning Boosting Algorithms to Detect the Fake User Profiles on Instagram," *iq-CCHES 2023 - 2023 IEEE Int. Conf. Quantum Technol. Commun. Comput. Hardw. Embed. Syst. Secur.*, pp. 1–7, 2023, doi: 10.1109/iq-CCHES56596.2023.10391800.
- A. R. Lubis et al., "Deep neural networks approach with transfer learning to detect fake accounts social media on Twitter," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 33, no. 1, pp. 269–277, 2024, doi: 10.11591/ijeecs.v33.i1.pp269-277.
- F. Benabbou, H. Boukhouima, and N. Sael, "Fake accounts detection system based on bidirectional gated recurrent unit neural network," *Int. J. Electr. Comput. Eng.*, vol. 12, no. 3, pp. 3129–3137, 2022, doi: 10.11591/ijece.v12i3.pp3129-3137.
- D. Amankeldin, L. Kurmangazyeva, A. Mailybayeva, N. Glazyrina, A. Zhumadillayeva, and N. Karasheva, "Deep Neural Network for Detecting Fake Profiles in Social Networks," *Comput. Syst. Sci. Eng.*, vol. 47, no. 1, pp. 1091–1108, 2023, doi: 10.32604/csse.2023.039503.
- R. Bhosale and V. Mane, "Enhancing User Trust: A Novel Hybrid Model to Detect Fake Profiles in Online Social Networks," *Intell. Syst. Appl. Eng. Eng.*, 2024, [Online]. Available: [www.ijisae.org](http://www.ijisae.org)
- M. VAROL ARISOY and T. TUNÇ ABUBAKAR, "A Comparative Analysis of Ensemble Learning Methods on Social Media Account Detection," *J. Eng. Technol. Appl. Sci.*, vol. 8, no. 2, pp. 87–105, 2023, doi: 10.30931/jetas.1325483.
- K. Zarei, R. Farahbakhsh, N. Crespi, and G. Tyson, "Impersonation on Social Media: A Deep Neural Approach to Identify Ingenuine Content," *Proc. 2020 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Mining, ASONAM 2020*, pp. 11–15, 2020, doi: 10.1109/ASONAM49781.2020.9381437.
- B. Yashwanth and K. Jaisharma, "An Automated Users Profile Classification in Instagram to Improve the Precision Using an Integrated Neural Network Model and Decision Tree Algorithm," *Proc. Int. Conf. Contemp. Comput. Informatics, IC3I 2023*, vol. 6, pp. 2553–2557, 2023, doi: 10.1109/IC3I59117.2023.10397941.
- S. Pan and T. Sharma, "Identifying Fake Instagram Profiles Using Machine Learning Techniques: An Empirical Analysis," *Proc. 2025 4th OPJU Int. Technol. Conf. (OTCON) Smart Comput. Innov. Adv. Ind. 5.0*, 2025, pp. 1–5, doi: 10.1109/OTCON65728.2025.11069428.
- Plikynas, D., Rizgelienė, I., and Korvel, G., "Systematic Review of Fake News, Propaganda, and Disinformation: Examining Authors, Content, and Social Impact Through Machine Learning," *IEEE Access*, vol. 13, 2025, pp. 17583–17624, doi: 10.1109/ACCESS.2025.3530688.