Mesopotamian journal of Cybersecurity
Vol. 5, No.2, **pp**. 804–821
DOI: https://doi.org/10.58496/MJCS/2025/046; ISSN: 2958-6542
https://mesopotamian.press/journals/index.php/cybersecurity

Research Article

An Artificial Intelligence-based System for Detecting Meta Fake Profiles via Gradient Boosting and Multilayer Perception

Ruaa A. Al-Falluji[1*], Marwan Ali Albahar [2], Ahmad Mousa Altamimi [3]

[1] *Department of Automation and Artificial Intelligence Engineering, College of Information Engineering, Al-Nahrain University, Baghdad, Iraq*

[2] *Department of Computing, College of Engineering and Computing in Al-Lith, Umm Al-Qura University, Makkah, Saudi Arabia*

[3] *Princess Sumaya University for Technology, King Hussein School of Computing Sciences, Amman, Jordan*

ARTICLEINFO

Article History

Received 25Apr 2025
Revised 6 May 2025
Accepted 23 Jul 2025
Published 10 Aug 2025

Keywords

Artificial intelligence
Classification
Explainable AI
Fake profile detection,
Gradient-boosted trees
LIME
Machine learning
Metaverse
Model interpretability
Multilayer perceptron
Pattern recognition
SHAP

ABSTRACT

The rise of metaverse platforms has renewed interest in detecting fake profiles, which pose a significant threat to digital ownership and asset transactions within these virtual environments. If digital ownership is not guaranteed, platforms risk missing the point of the metaverse. Current supervised learning techniques for fake profile detection often struggle to maintain acceptable accuracy and interpretability in practice. To address this problem, this study investigates the application of two machine learning models, multilayer perceptron and gradient boosted trees, for detecting fake profiles, with model evaluation performed via two Explainable AI (XAI) techniques, Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP). The algorithms were implemented and evaluated on a dataset of 1244 profiles (1043 real and 201 fake) with 12 attributes. The major finding is that both techniques perform well, but the gradient boosting model achieves a higher accuracy of 99.2% compared with the multilayer perceptron's accuracy of 98.4%. Furthermore, the LIME and SHAP analyses provide insights into the feature importance and decision-making processes of the models. These results suggest that gradient boosting, in conjunction with explainable AI methods, is a more accurate, interpretable, and representative solution for detecting fake profiles in real-world metaverse scenarios, contributing to the development of more secure and reliable digital ownership within these platforms.

1. INTRODUCTION

Facebook, a social media giant with three of the largest social media apps (Facebook, Instagram, and WhatsApp), was renamed Meta Platforms [1]. Meta reflects the company's growing ambitions beyond social media to embrace Facebook's spirit of creativity. Meta is helping to build the metaverse, the next evolution of social connection that provides a 3D digital space that allows its participants to augment reality with digital simulations to meet, work, or collaborate as realistically as possible [2]. Participants can create virtual office spaces, attend concerts, travel, or even shop for virtual clothes and goods. For example, the value of real estate circulation in the metaverse is equivalent to $2.4 million last in November 2021 [3]. The possibilities are as broad as imagination. Nevertheless, if Meta does not enable digital ownership, it may undermine the very essence of the metaverse. To that end, the company shut down 1.3 billion fake accounts on its main platform between October and December 2021, yet millions are likely to remain [4]. More than 12 million pieces of content from fake profiles about COVID-19 and vaccines were removed after being flagged by global health experts as misinformation [5]. Problems commonly caused by fake profiles include privacy violations, impersonation of a specific identity, or misuse, such as spreading hate speech and cyberbullying [6]. These

*Corresponding author. Email: Ruaa.adeeb@nahrainuniv.edu.iq

threats can be categorized as cyber impersonation, social engineering attacks, identity theft, misinformation campaigns, and online harassment, which pose serious risks to digital trust and personal safety. While content-removal algorithms exist, the challenge of effectively identifying and mitigating fake profiles remains [7]. Existing machine learning (ML) models offer promise in automating fake profile detection through profile feature analysis. However, many of these models operate as "black boxes," providing limited insight into the decision-making process. This lack of transparency affects user trust, limits the ability to identify and correct errors, and impedes our understanding of the evolving tactics employed by fake profile creators. This study seeks to answer the following key research questions:

• Can machine learning models accurately detect fake profiles on the basis of structured profile features? • How can explainable AI (XAI) methods enhance the transparency and interpretability of such models? • Which model, gradient boosting or multilayer perceptron, performs better in this classification task?

In this paper, we present a machine learning-based model designed to automatically identify fake profiles by analysing structured profile features. We explore two supervised learning techniques, gradient boosting and multilayer perception, and integrate them with two explainable AI methods, SHAP and LIME, to achieve transparency and trust in the model outcomes. Our approach was implemented via Jupyter Notebook and a dataset of 1244 profiles (1043 real and 201 fake). Each profile includes 12 attributes, seven non-behavioral attributes which are related to the profile content (profile picture, workplace, education, living place, check-in, introduction, and family relationships). The five remaining attributes are numerical behavioral (number of liked pages, number of groups, number of posts, number of tags, and number of mutual friends). Two experiments were conducted with and without missing values, and the results revealed minimal performance differences due to the low number of missing values. Gradient boosting achieved 99.2% accuracy, outperforming the multilayer perceptron (98.4%). Our key contributions include the following:

• We propose an XAI-augmented detection architecture that combines a high-accuracy classifier with two interpretability layers that include SHAP for feature attribution and LIME for visual explanations of both global and local aspects to transform the model into a transparent decision aid that reveals how individual profile signals lead to fraud verdicts.
• We compare gradient-boosting and multilayer perceptron baselines through an evaluation to measure both predictive accuracy and interpretability fidelity, which shows that the integrated XAI pipeline achieves state-of-the-art accuracy while providing social media operators, regulators and end-users with actionable insights for large-scale manipulation and misinformation and identity fraud prevention.
The rest of this paper is organized as follows: Section 2 discusses the related work proposed to detect fake profiles. The research methodology is described in Section 3. This includes the dataset, the acquisition process, the missing values, and the utilized profiles' attributes. Section 4 presents the proposed model, while the experimental results, model evaluation via the LIME and SHAP methods and a comparison with previous results are given in Section 5. The limitations of this study are presented in Section 6. Finally, the study conclusions and future work are given in Section 7.

2. RELATED WORK

With the rapid spread of online social networks, the fake profile phenomenon is becoming a significant concern [8]. Different approaches have been proposed in the literature to detect fake profiles in Meta, Twitter, and LinkedIn. Machine learning techniques dominated these works with different analyses of view. For example, supervised algorithms (e.g., support vector machines, decision trees, and naïve Bayes) were implemented to detect fake profiles in Meta [9]. The authors collected their dataset (975 profiles), where attributes such as education, work, and gender were utilized in the analysis.

In the same vein, the authors of [10] collected their dataset from the Meta API and considered 17 behavioral attributes (e.g., shares, comments, tags, and alike attributes), as they are enough to reflect the user's interaction with other profiles. The authors implemented 12 supervised machine learning techniques. However, the results obtained were not promising. On the other hand, the work in [11] analysed the visit histories of the passive profile and the links between that profile and active comments. Then, a specific interaction graph is created to describe the users' behavior and identify fake accounts. A recent study [12] considered a dataset of 982 profiles (781 real and 201 fake) with a set of 12 different attributes (behavioral and nonbehavioral). The dataset was collected by developing a custom web crawler, which automatically accessed and extracted publicly available profile information from the target platform. The collected dataset was examined via a set of supervised and unsupervised techniques. The obtained results revealed that the supervised algorithms achieved higher accuracy than unsupervised techniques did.

*Corresponding author. Email: Ruaa.adeeb@nahrainuniv.edu.iq

Although studies concerning meta-fraudulent account detection are rare due to the nonavailability of data and the difficulty of collection, categorization, expensive and time-consuming annotation[13], the work presented in [12] was the most promising and motivated us to carry out this study and consider more techniques to achieve better accuracy. With respect to the Twitter network, numerous machine learning techniques have been employed to detect fake profiles because of the ease of data collection. The authors of [14] utilized profile information to recognize fake profiles. In [15], a multivariable pattern-recognition approach was implemented using purchased fake accounts. According to their experiments, the authors found a strong relationship between the profile name, screen name, and email parameters of all fake accounts.

In [16] and [17], the authors utilized a method to recognize suspicious behavior. The method is based on the user's synchronization and abnormal activity. The proposed approach achieved high detection efficiency in detecting fake Twitter accounts. Other approaches have been suggested to identify fake Twitter accounts in the literature. The work of [18] was based on the analysis of features containing user profiles and tweets. The authors utilized a supervised machine learning technique to classify accounts as real or malicious on the basis of analysing and comparing text frequencies in the profiles' attributes. [19] identified fake Twitter accounts on the basis of the presence of geoinformation, the number of followers, and the availability of a hashtag in a tweet. In [20], the authors proposed a real-time approach for identifying fake accounts on the basis of threshold values that were selected according to the regular user features.

Ultimately, detecting fake profiles was also considered for the LinkedIn network. The authors of [21] employed two supervised classifiers (e.g., a neural network and support vector machine) to detect fake profiles on the basis of 15 static attributes (e.g., a summary of the profile and number of languages). The dataset used was collected manually, and it contains 74 profiles (40 legitimate and 34 fake). Their results show that the SVM achieves a higher detection efficiency than the NN does. In [22], the authors suggested an approach with three components: a distiller, a profile hunter, and a profile verifier. The distiller creates test queries and runs them on social networks and search engines. A user record is constructed on the basis of the results obtained against each query, which consists of a full user profile's name with user identifying terms. The output of the second component is employed to specify the potential LinkedIn profiles that belong to that user. The obtained results are collected to represent a profile record that links the real profile of a user to all returned potential profiles. The last component (profile verifier) checks the similarity with the user's real profile. The profiles are finally classified according to the similarity score. In [23], the authors used traditional machine learning and deep learning methods to detect fake profiles on online social networks. They employed datasets of varying sizes in their study. The results show that the synthetic minority oversampling technique (SMOTE) improves the performance on imbalanced data, with long short-term memory (LSTM) outperforming large datasets. In [24], the authors employed a model to address the threat of identity theft on social media. They analyzed social media features to distinguish between genuine and Sybil accounts. The suggested model also provides a visual representation of Sybil's presence in the user's social graph.

In [25], the authors used questionnaires to assess user awareness and perceptions of data breaches and analysed shared messages for linguistic patterns. Their study provided insights into how data breach risk influences user behavior and trust in digital platforms. As a trial to prevent suspicious behavior and protect user data on social platforms, the authors in [26] suggested a model to detect fake profiles on Facebook. The model uses a machine learning approach to analyse both public and private features. Multimedia big data, which includes diverse, large-scale content such as text, audio, and video, was used. In [27], the authors presented a systematic framework for detecting fraud on social platforms. It consists of three stages: component selection, feature extraction, and finally classification. Traditional and deep learning approaches were reviewed. The authors highlighted the main challenges, such as limited labelled data and bot-generated reviews. They also emphasized the need for multimodal data. Another study was developed in [28], in which the authors explored how machine learning and neural networks enhance social media content recommendation, emotion detection, and network management via advanced data analysis. They also highlighted the effects of these technologies, ethical challenges, and risks.

Machine learning algorithms were employed by the authors in [29] to analyse and classify profile information, network connections, and posting behavior. The classification of the accounts was implemented on the basis of that analysis and feature extraction process. In [30], the authors evaluated different machine learning algorithms to detect fake accounts. They focused on neural networks as the most common and effective approach. They suggested that integrating object detection approaches with machine learning methods can improve efficiency. The author of [31] discussed the rapid spread of disinformation, such as fake news and conspiracy theories, via the rise of social media, which affects trust in online platforms. It also emphasizes the importance of understanding disinformation and

*Corresponding author. Email: Ruaa.adeeb@nahrainuniv.edu.iq

enhancing defenses against it. The paper also highlights recent computational approaches that have been employed to detect disinformation. This also highlights the urgent need to strengthen resistance to online disinformation. In [32], the challenge of detecting misinformation on social media was addressed. A semi supervised learning framework was suggested. This framework uses real data to handle extreme class imbalances. It was tested on COVID-19-related Twitter data. The framework outperformed traditional techniques such as SMOTE and generative adversarial networks (GANs). The research community has presented many solutions to the fake profile detection problem on online social networks (OSNs). Numerous studies have employed various machine learning methods with specific analyses of those profiles. Notably, few studies have considered meta profiles, and to the best of our knowledge, no one has implemented gradient boosting and multilayer perception with explainable AI, as in our work. A comparative analysis of the common machine learning-based approaches related to the detection of fake profiles in online social networks is shown in Table I.

TABLE I: COMPARATIVE ANALYSIS OF MACHINE LEARNING APPROACHES FOR FAKE PROFILE DETECTION IN ONLINE SOCIAL NETWORKS

No	Author Name	Methods	Limitations	Accuracy
1	M. Albayati and A. Altamimi ([12])	Supervised and Unsupervised ML on 982 profiles	Relied on a manually developed crawler. Dataset size is limited, and unsupervised models performed poorly compared to supervised ones.	96%
2	R. V. Kotawadekar, A. S. Kamble, and S. A. Surve ([9])	Supervised ML (SVM, Decision Tree, naïve Bayes)	Used a small dataset (975 profiles), limiting generalization. Attributes used were mainly static (education, work, gender), potentially omitting behavioral nuances.	94%
3	S. Shehnepoor, R. Togneri, W. Liu, and M. Bennamoun ([27])	Systematic framework (feature extraction, classification)	Challenges include lack of labelled data, difficulty in using multimodal data, and evasion tactics.	94%
4	A. Shah, S. Varshney, and M. Mehrotra ([23])	ML and Deep Learning with SMOTE	Performance is highly dependent on data size and quality. SMOTE may introduce synthetic noise.	93%
5	G. Kontaxis, I. Polakis, S. Ioannidis, and E. P. Markatos ([22])	Three-component system: distiller, hunter, verifier	The system is complex, resource-intensive, and assumes available web data for profile verification.	92%
6	C. Xiao, D. Freeman, and T. H. P. ([18])	Supervised ML on tweets and profile features	Only textual and static profile features were used. Lacks deep semantic or contextual behavioral features.	91%
7	J. Ezarfelix, N. Jeffrey, and N. Sari ([30])	Neural networks and object detection	Integration complexity requires high computational resources. Object detection is rarely applicable in text-dominant platforms.	91%
8	S. Adikari and K. Dutta ([21])	SVM and Neural Network on LinkedIn	Very small dataset (74 profiles) and static features limit generalizability. The manual collection introduces bias.	90%
9	S. R. Sahoo and B. B. Gupta ([26])	ML on multimedia big data (text, audio, video)	Requires large, diverse datasets and complex processing pipelines. Privacy and scalability concerns.	90%
10	K. Thomas, D. McCoy, C. Grier, A. Kolcz, and V. Paxson ([15])	Multivariable pattern recognition	Training on purchased fake accounts may not generalize well to organically occurring fakes. Possible overfitting to specific patterns.	89%
11	H. Ali, I. Malik, S. Mahmood, F. Akif, and J. Amin ([24])	Model to detect Sybil accounts	Targeted only Sybil-type attacks. May not detect non-Sybil fake accounts effectively.	89%
12	G. Stringhini, C. Kruegel, and G. Vigna ([14])	Profile-based feature analysis on Twitter	Used limited profile-based features, which may be easily manipulated by fake accounts. Ignores dynamic user behavior.	88%
13	M. Chakraborty, S. Das, and R. Mamidi ([29])	ML analysis of profile info, connections, and posts	Model performance depends heavily on feature engineering and availability of connection data.	88%
14	A. Elazab, A. M. Idrees, M. A. Mahmoud, and H. Hefny ([19])	Geo-information, follower count, hashtags	Feature set is simple and easy to spoof. No behavioral pattern analysis.	87%

Table I shows that machine learning approaches for fake profile detection often suffer from limitations such as small or static datasets, which impact the generalizability of the models. Studies with high accuracy scores may be inflated due

*Corresponding author. Email: Ruaa.adeeb@nahrainuniv.edu.iq

to the limited variation in these datasets. Deep learning models show potential but face challenges in terms of computational costs and data requirements. Integrating behavioral and multimodal features is crucial for effective detection, and addressing the lack of interpretability remains a key area for improvement.

3. RESEARCH METHODOLOGY

This work considers detecting fake profiles in Meta with high accuracy. To this end, we propose a model based on gradient boosting and multilayer perception techniques. The model is implemented via Jupyter Notebook with a dataset containing 1244 profiles (1043 real and 201 fake) adopted from the work presented in [12]. The reason behind adopting the same dataset is the strict privacy setting imposed by the Meta company. Each profile consists of 12 attributes and is divided into 5 behavioral and 7 nonbehavioral. These attributes are discussed in the next section. Before conducting the experiments, we noted that the dataset contained missing values. Specifically, there are 73 missing values from the number of likes attributes and 61 missing values from the number of group attributes. Thus, we run our model twice, first replacing missing values with zero values and then replacing missing values with the average of their respective classes. The analysis of the results in both cases revealed that the results are almost the same.

After handling the missing values, we trained our model. Sixty percent of the total dataset is used as a training dataset, 20% of the dataset is used as a validation set, and the remaining dataset is utilized for testing. After that, we run our model. In the case of a multilayer perception classifier, three hidden layer sizes of 10, 50, and 100 are used. For the activation function, ReLU, tanh, and logistics are used. For the learning rate, constant, scaling, and adaptive methods are used. After running the model, the best multilayer perception estimator had a ‘tanh’ activation function, a hidden layer size of 50, and an adaptive learning rate. In the case of gradient boosting classifiers, 5, 50, and 250 are used as the n_estimator, and 1, 3, 5, and 7 are used as the maximum depth, whereas 0.01, 0.1, 1, and 10 are used as the learning rate. After running the model, the best gradient boosting classifier has 250 as the n_estimator, 1 as the max depth, and 0.1 as the learning rate.

4. PROPOSED MODEL

In this section, we discuss the model. However, before that, the dataset used will be described.

4.1 Dataset

The employed dataset contains 1244 profiles. Each profile consists of 12 attributes that are divided into nonbehavioral and behavioral attributes. A brief description of each attribute is given next, and more illustration is shown in Table II.

The nonbehavioral attributes (profile content attributes) include the following:

- Profile Picture: This attribute is used as an indicator for real profiles, where a real profile picture is represented by "1" and "0" represents a fake profile picture.
- Workplace: This attribute contains information about the user's workplace. "1" is used for an available workplace and "0" for a not-of-available workplace.
- Education: This attribute refers to the last educational institution in which that user enrolled. A valid existing institute name is represented by "1", and "0" represents an absent or inaccurate institute.
- Living Place: This indicates the living address of the user. A value of "1" refers to a validly mentioned living place, whereas "0" represents an invalid or absent living place.
- Check-In: This identifies the places that the user visited and announces the visit. "1" is used even if the user registered one check-in, and "0" represents no check-in registration.
- Introduction Bio: A valid mentioned biography of a user is represented by "1" (even if a user wrote at least five words as a biography), and "0" is used to describe an invalid attribute.
- Family/Relationship: This refers to the status of a user's social relationship. "1" is used to represent a valid status, and "0" represents an invalid or absent status.

*Corresponding author. Email: Ruaa.adeeb@nahrainuniv.edu.iq

TABLE II: DATASET ATTRIBUTES

Attribute Name	Type	Justification
Profile Picture [12]	Non-Behavioral/Profile Content Attribute	Legitimate users often put their real pictures compared to fake users.
Education [33]	Non-Behavioral/Profile Content Attribute	Legitimate users often mention their real information about education in their profiles rather than fake users.
Workplace [33]	Non-Behavioral/Profile Content Attribute	Legitimate users often put their real workplace information in their profiles rather than fake users.
Living Place [12]	Non-Behavioral/Profile Content Attribute	Legitimate users often use their real information about living places in their profiles rather than fake users.
Check-In [12]	Non-Behavioral/Profile Content Attribute	Real users often check in the visited places than fake users.
Introduction "Bio" [33]	Non-Behavioral/Profile Content Attribute	Legitimate users often write their biographies compared to fake users.
Family/Relationship [12]	Non-Behavioral/Profile Content Attribute	Legitimate users often share their social status and relationships compared to fake users.
Number of Pages [20]	Behavioral/Numerical Attribute	Real users often like more pages than fraudulent users
Number of Groups [12]	Behavioral/Numerical Attribute	Actual users often participate in more groups than fraudulent users.
Number of Posts [33][11]	Behavioral/Numerical Attribute	Real users often have more posts such as text posts, videos, links, etc., than fraudulent users.
Number of Tags [33][34]	Behavioral/Numerical Attribute	Legitimate users are more tagged by other users than fraudulent users.
Number of Mutual Friends [35]	Behavioral/Numerical Attribute	Legitimate users usually have more mutual friends with the tested profile than fraudulent users.

Table II outlines the attributes used for fake profile detection, categorizing them as either nonbehavioral (profile content) or behavioral/numerical. The justification column explains why each attribute is indicative of either a legitimate or fake user. Profile content attributes, such as profile picture, education, and workplace, are often genuine for real users but may be falsified in fake profiles. Behavioral attributes, such as the number of pages liked, groups joined, posts, tags, and mutual friends, reflect a user's activity and connections, with legitimate users typically exhibiting greater engagement than do fake users.

The behavioral attributes (numerical attributes) include the following:

- Number of Pages: This attribute reflects the number of pages that are liked by a user. It is represented by a natural number (N).
- Number of Groups: The natural (N) number represents the number of groups in which a user is enrolled.
- Number of Posts: This attribute reflects the number of posts that a user posted on his/her timeline. These posts include photos, text posts, videos, etc. A natural number (N) is used to represent this attribute.
- Number of Tags: This refers to the number of links or tags that other users post on a tested user timeline, which is represented by the (N) natural number.
- The number of mutual friends: This variable represents the number (N natural number) of mutual friends between the tested profile and the model's user.

4.2 Classification Model

Two supervised learning methods were used in this study to detect fake profiles on the basis of a set of 12 different attributes. Among these attributes, 11 attributes are independent (e.g., profile picture, no. of pages, no. of groups, no. of mutual friends, education, workplace, living place, relationship, check-in, no. of posts, no. of tags and Intro), and one

*Corresponding author. Email: Ruaa.adeeb@nahrainuniv.edu.iq

is dependent (class). The multilayer perception and gradient boosting techniques are used to train the system to classify the profiles. In the multilayer perception process, the attributes are fed to multilayer perception. Supervised learning is implemented via a nonlinear activation function in the hidden layer. The output is either real or fake.

In the gradient boosting technique, data are first assigned equal weights, the model trains, and the weight is increased at the other level to improve our training dataset. Here, a decision tree is used to construct a strong decision tree for training the model to provide the best results with the highest accuracy. In other words, within the gradient boosting technique, the individual models are made by placing more weight on instances with incorrect predictions and high errors. In the model, more focus is given to cases that are hard to predict so that the model learns from past mistakes, known as a loss function. The gradient is used to minimize the loss function. In each training step, weak learners are built, and their predictions are compared with the actual predictions. The model aims to reduce the error between the actual and predicted outcomes so that the model with the lowest error rate can be built to achieve the best accuracy.

4.3 Model architecture and mathematical representations

The mathematical representation of the model is illustrated below:

Step 1: Input representation: $X \in R^{N \times d}$, $y \in \{0,1\}$ N Eq. (1)

where the data matrix $X \in R^{N \times d}$ contains *N* user profiles, each represented by *d* engineered attributes, while the binary label vector $y \in \{0,1\}^N$ encodes the ground-truth identity (1 = fake, 0 = genuine). This notation fixes the supervised learning setting and provides the basis for subsequent model definitions.

Step 2: Model 1 (MLP): $\hat{y} = \sigma\left(W^{(2)}\phi\left(W^{(1)}x + b^{(1)}\right) + b^{(2)}\right)$ Eq. (2)

specifies the *forward propagation* of the multilayer perceptron (MLP). An input vector *x* is first linearly transformed by the weight matrix $W^{(1)}$ and bias $b^{(1)}$ passed through a nonlinear activation $\phi(\cdot)$ and then mapped by a second affine layer ($W^{(2)}$, $b^{(2)}$). The final logistic function $\sigma(\cdot)=1/(1+e^{-z})$ converts the latent score to a probability estimate $\hat{y} \in [0,1]$. In particular, it quantifies the divergence between the predicted and true labels and is minimized during back-propagation to fit the network parameters (Eq. 3).

$$MLP = -\frac{1}{n}\sum_{i=1}^{N}[y_i \log(\hat{y}y_i) + (1 - y_i)\log(1 - \hat{y}y_i]$$ Eq. (3)

Step 3: Model 2 (gradient boosting): $F_0(x) = argmin_\gamma \sum_i \tau(y_i, \gamma)$ Eq. (4)

initiates *gradient boosting* by selecting a constant function $F_0(x)$ that minimizes the chosen loss $\tau(y_i, \gamma)$ over all training samples, typically the negative loglikelihood for binary classification. This initialization anchors the additive boosting sequence.

Equation (5) presents the *stagewise update*: at iteration *m*, the current ensemble prediction $F_{m-1}(x)$ is augmented by a newly fitted weak learner $h_m(x)$ scaled by the learning rate $\nu \in [0,1]$. The procedure iteratively refines the model in the direction of the loss gradient.

$$F_m(x) = F_{m-1}(x) + \nu h_m(x)$$ Eq. (5)

Equation (6) gives the *logistic loss* adopted here: where $F^{(xi)}$ is the aggregated score after the final boosting round. The formulation translates the binary classification task into a convex optimization problem amenable to gradient-based boosting.

$$L_{GB} = \sum_i log(1 + e^{-yiF^{(xi)}})$$ Eq. (6)

Step 4: Evaluation (Cross Validation)

Equation (7) defines the *K-fold cross-validation accuracy*. The dataset is partitioned into *K* disjoint folds D_J ; for each fold, predictions $\widehat{Y1}$ generated by a model trained on the remaining K−1 folds are compared with the held-out labels Y_I . The indicator function $1[\widehat{Y1} = Y_I]$ equals 1 when a prediction is correct. Averaging first over instances in fold *J* and then over all folds yields an unbiased estimate of out-of-sample accuracy.

$$CV - Score = \frac{1}{K}\sum_{J=1}^{K}\frac{1}{|D_J|}\sum_{I \in D_J} 1[\hat{Y}Y_I = Y_I]$$ Eq. (7)

Step 5: Local Explanation (LIME)

*Corresponding author. Email: Ruaa.adeeb@nahrainuniv.edu.iq

Equation (8) introduces *Local Interpretable Model-agnostic Explanations (LIME)*. Around a specific test point, the complex predictor is approximated by an interpretable linear surrogate

$$g(x) = \phi_0 + \sum_{J=1}^{D} \beta_J X_J \text{ (approximate model) Eq. (8)}$$

where $D \leq d$ denotes the subset of perturbed features deemed locally important, β_J represents the locally fitted coefficients, and ϕ_0 represents the intercept. The sparsity and linearity make the contribution of each selected feature transparent to the end user.

Step 6: Global + local explanation (SHAP):

Equation (9) formalizes the *SHapley Additive exPlanations (SHAP)* decomposition:

$$f(x) = \phi_0 + \sum_{J=1}^{D} \phi_0 \; X_J \text{ Eq. (9)}$$

Here, $f(x)$ is the original model's output, $\phi_0 == Ex\,[f(x)]$ is the baseline (expected prediction over the data distribution), and each ϕ_J is a Shapley value that allocates the difference $f(x) - \phi_0$ fairly among features according to cooperative game theory. Unlike LIME, SHAP provides both local fidelity and a globally consistent attribution scheme.

The detailed architecture of the suggested model is shown in Figure 1:

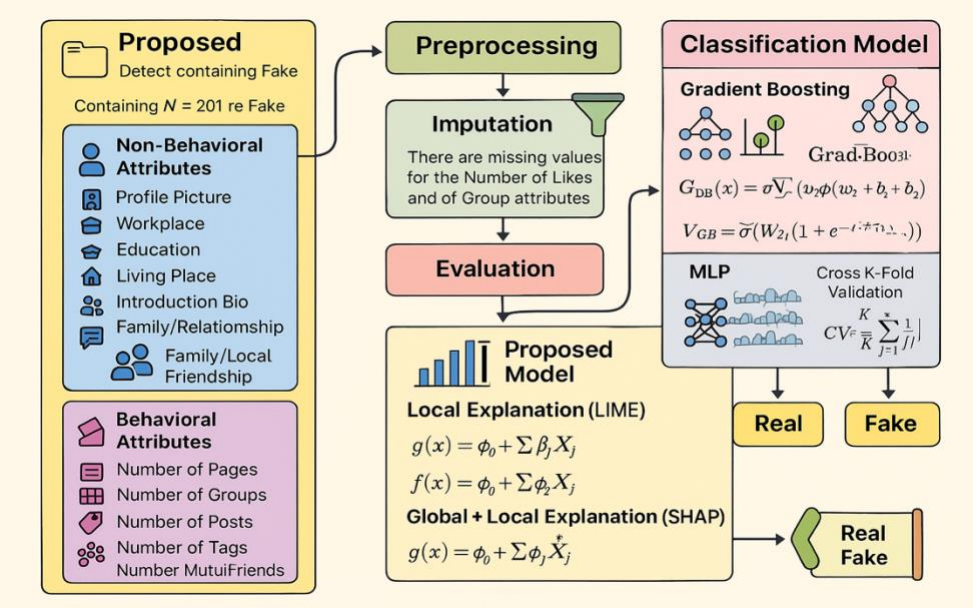

Fig. 1. Fig. 1: Proposed architecture model

5. EXPERIMENTAL RESULTS

In this section, a discussion of the evaluation measurements and evaluation process is presented.

5.1 Performance Measurements

To evaluate any classification technique, metrics that measure the overall performance are used: accuracy, precision, recall, and specificity. These metrics are joined directly with the confusion matrix.[36]. The numbers of correct and incorrect predictions for each class are tabulated in the confusion matrix. This matrix shows where the suggested model is confused when making predictions. The results of multilayer perception and gradient boosting are recorded in Tables III and IV, respectively. Both models demonstrate high accuracy, with GB achieving a slightly higher accuracy of 99.2% compared with the 98.4% accuracy of MLP. While both models exhibit strong precision and recall for real profiles, GB shows superior specificity (0.850) compared with MLP (0.704), indicating a better ability to correctly identify fake profiles. This suggests that GB is more effective in minimizing false positives when detecting fake profiles. Figure 2 shows the detailed evaluation metrics for multilayer perception and gradient boosting.

*Corresponding author. Email: Ruaa.adeeb@nahrainuniv.edu.iq

TABLE III: CONFUSION MATRIX OF MULTILAYER PERCEPTION

Actual Status		**Real**	**Fake**
Predicted profiles	**Real**	984	59
	Fake	59	142
Accuracy	**Precision**	**Recall**	**Specificity**
0.984	0.943	0.943	0.704

TABLE IV: GRADIENT BOOSTING CONFUSION MATRIX

Actual Status		**Real**	**Fake**
Predicted profiles	**Real**	1013	30
	Fake	30	170
Accuracy	**Precision**	**Recall**	**Specificity**
0.992	0.971	0.971	0.850

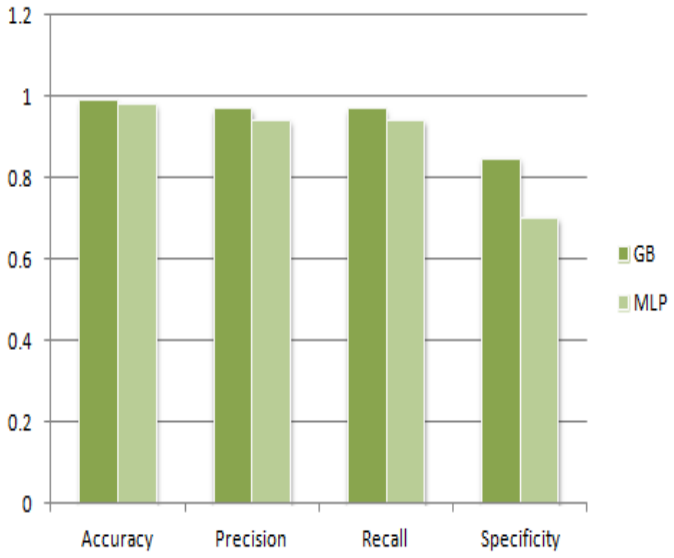

Fig. 2. Performance Measurements for Multilayer Perception and Gradient Boosting

5.2 Data analysis

The graphical analysis of the model gives us an educative lead in deciding whether the profile is fake or real. Our data contain Boolean as well as numeric values, so we perform graphical representation for both types of attributes. The graphical analysis of nonbehavioral attributes is shown in Figures 3 to 8.

*Corresponding author. Email: Ruaa.adeeb@nahrainuniv.edu.iq

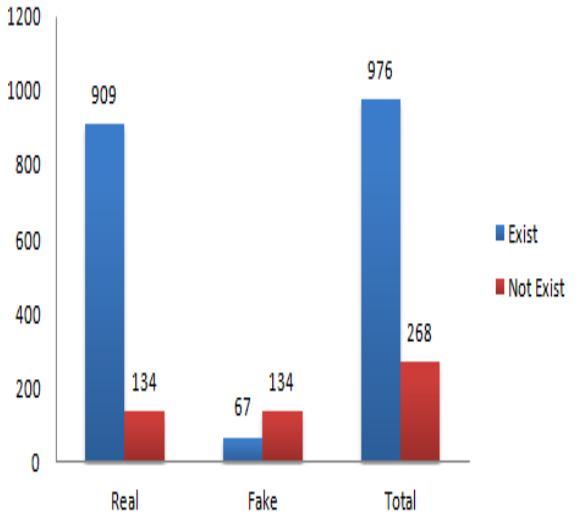

Fig. 3 Living Place

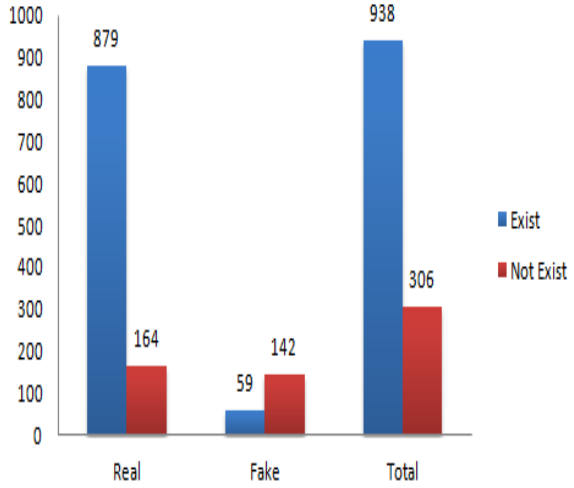

Fig. 4 Education

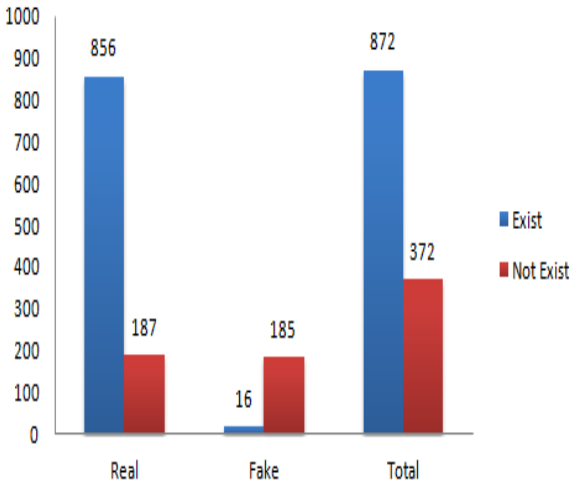

Fig. 5 Check In

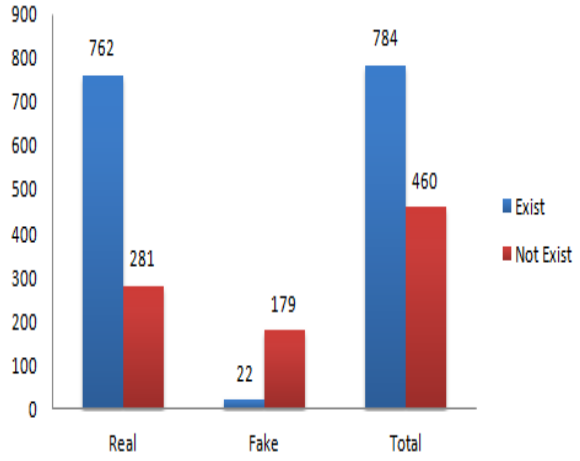

Fig. 6. Family

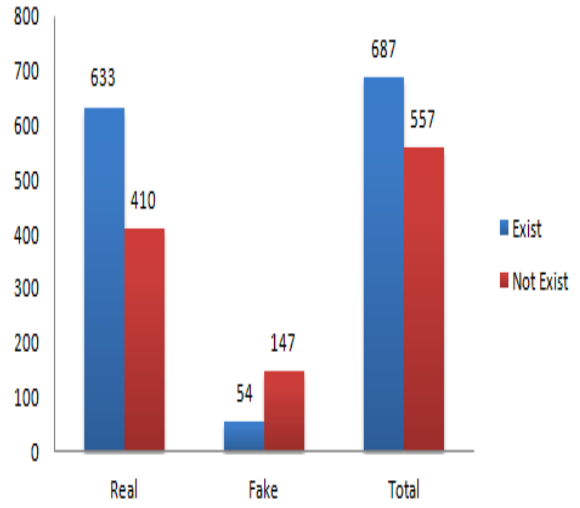

Fig. 7 Work

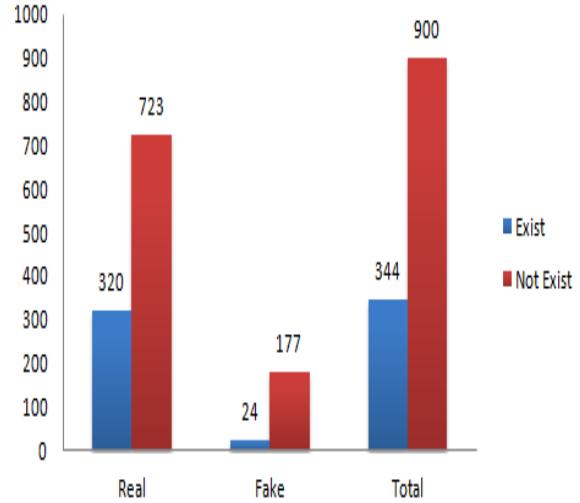

Fig. 8. Intro

*Corresponding author. Email: Ruaa.adeeb@nahrainuniv.edu.iq

For the attributes with Boolean values, 87.15% of the real profiles are related to living places, 84.27% are related to education, and 82.07% are related to check-in, which are substantial percentages that can be used as indicators for real profiles. With respect to the family/relationship attribute, 73.05% of the real profiles have details, but in the case of a fake profile, 88.17% of the details do not exist. Real profiles have 60.69% details regarding the workplace, but in the case of a fake profile, 72.14% of details do not exist. With respect to the intro attribute, 30.69% of the real profiles have details, but with respect to the fake profile, 87.19% of the details do not exist. Therefore, these details can be used to determine fake profiles, as the percentage of absence is greater in this case. The analysis of the numerical attributes is represented in Figures 9 to 13.

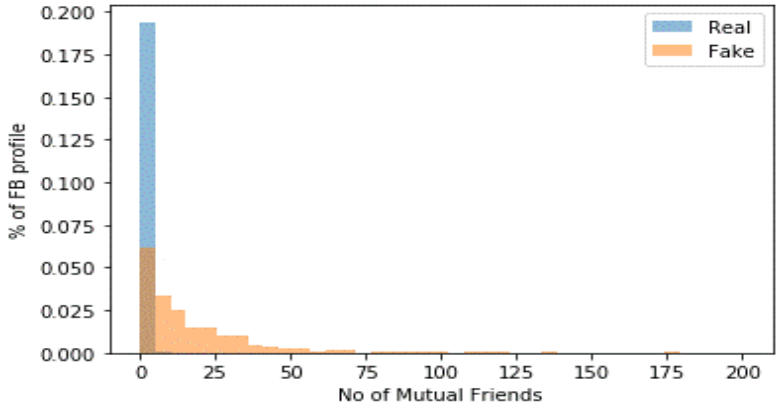

Fig. 9. Mutual friends

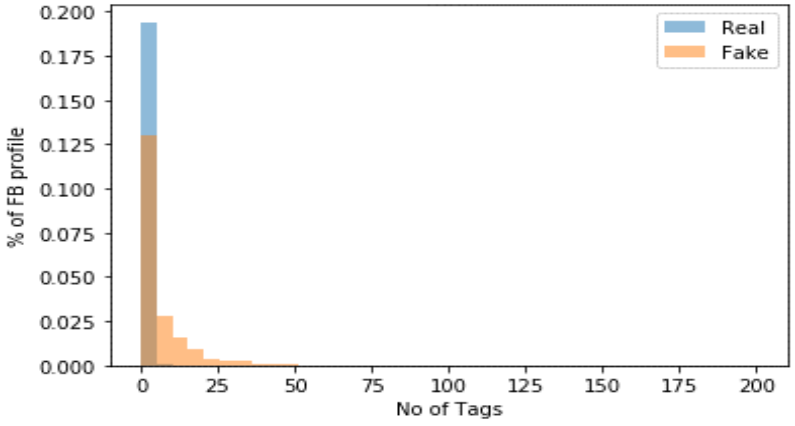

Fig. 10 Tags

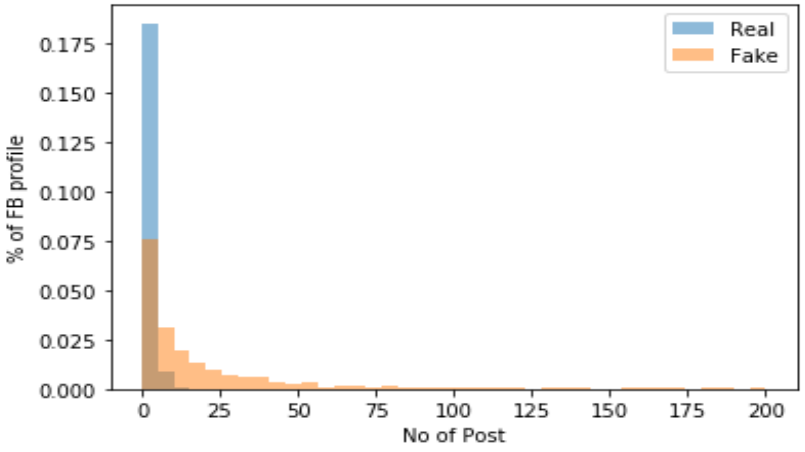

Fig. 11. Posts

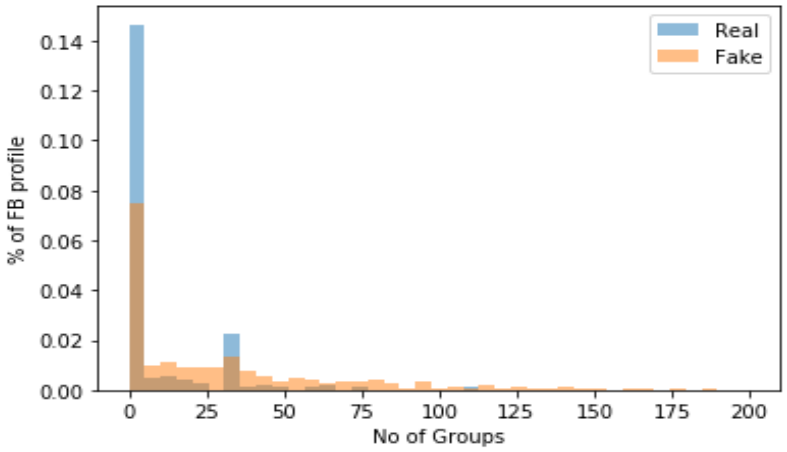

Fig. 12. Groups

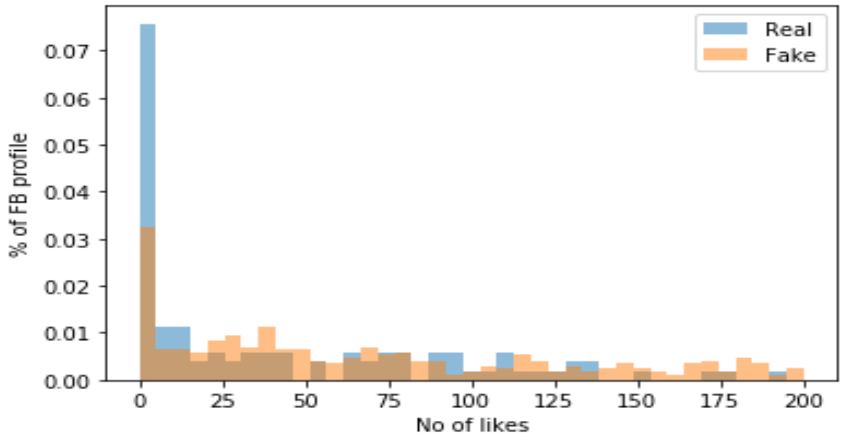

Fig. 13. Likes

Finally, we compared the results of gradient boosting and multilayer perception with the results obtained by the authors in [12] because they used the same dataset but with different supervised and unsupervised classifiers. For a fair comparison, we compared their results in the case of handling the missing values problem, which they used the K-NN estimator to solve. The comparison is shown in Table V. GB achieves the highest accuracy (0.992) among the compared algorithms. It is clear that it is excellent in working on unprocessed data and data with missing values. While ID3 has competitive precision and recall, its specificity is lower than that of GB. The k-means and k-medoids algorithms demonstrate significantly lower performance across all the metrics, highlighting the superiority of our models and ID3 for fake profile detection. The MLP model also outperforms the SVM and k-NN models in terms of accuracy.

TABLE V. COMPARISON WITH OTHER TECHNIQUES LISTED IN [12]

Algorithm	**Accuracy**	**Precision**	**Recall**	**Specificity**
ID3	0.977	0.987	0.984	0.950
SVM	0.957	0.978	0.968	0.915
k-NN	0.914	0.952	0.939	0.815
k-Means	0.673	0.766	0.846	0.000
k-Medoids	0.670	0.884	0.673	0.656
MLP (our model)	***0.984***	***0.943***	***0.943***	***0.704***
GB (our model)	***0.992***	***0.971***	***0.971***	***0.850***

5.3 Model Validation via K-Fold

Model validation plays a vital role in checking model performance, and determining whether a model performs effectively without overfitting. For this purpose, we tested both models via cross-K-fold validation with K-fold = 5, K-fold = 10, and K-fold = 20. Upon testing the models, we obtain the following results:

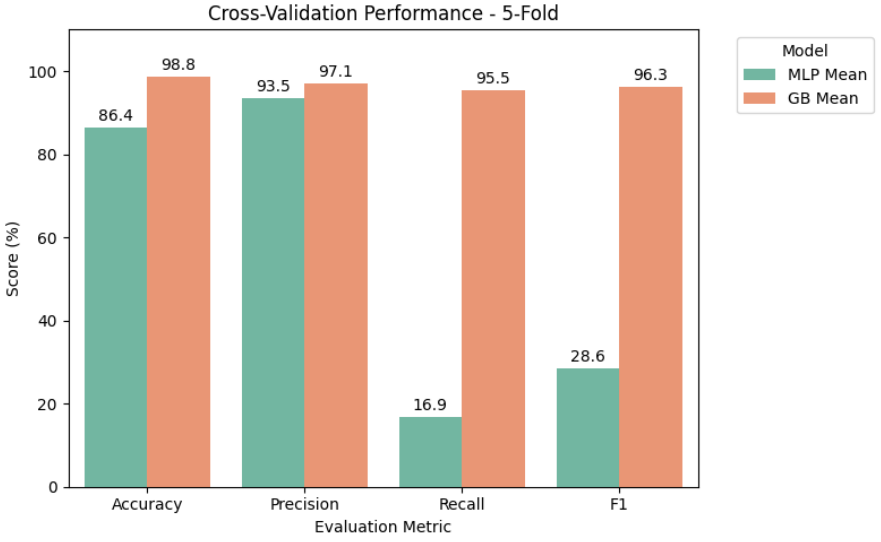

Fig. 2. Fig. 14: 5-fold validation for both models (MLP and GB)

The results show that gradient boosting performed well and achieved the highest accuracy compared with the multilayer perceptron model. A comparison of all the parameters reveals that the GB method performs well in fake profile classification.

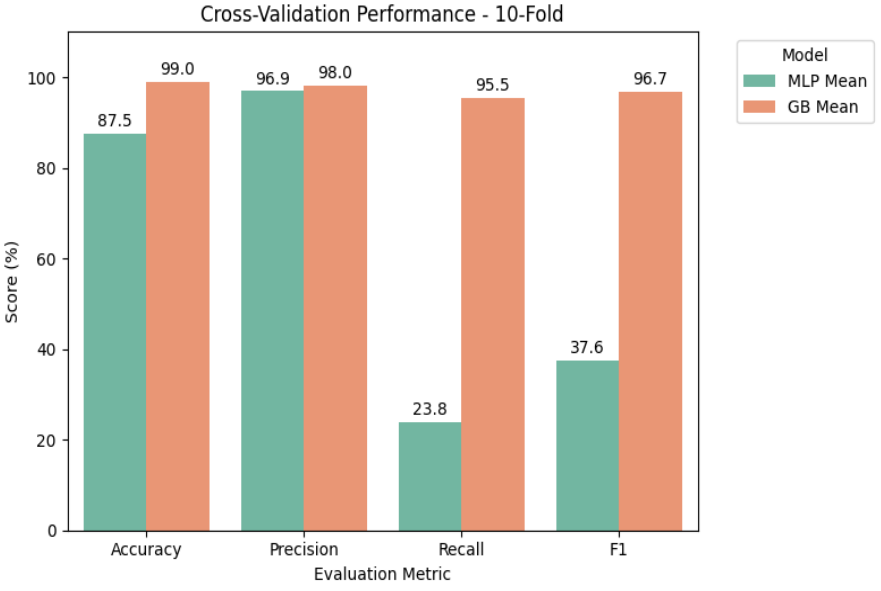

Fig. 3. Fig. 15: 10-fold validation for both models (MLP and GB)

Similarly, the 10-fold validation again revealed the superior performance of gradient boosting over multilayer perceptron by achieving 99% accuracy with 98% precision, 95.5% recall and 96.7% F1 scores compared with MLP, which had lower performance in terms of recall and F1. This alternatively means that the MLP performance is affected by overfitting.

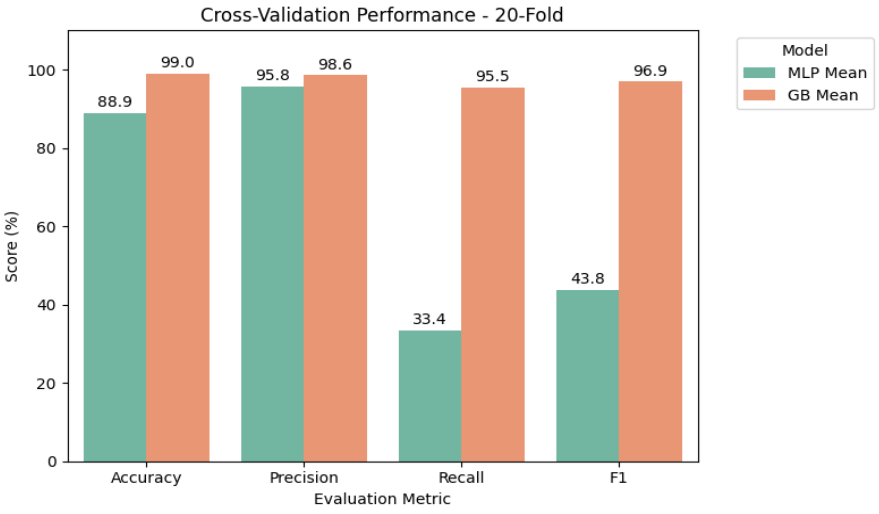

Fig. 4. Fig. 16: 20-fold validation for both models (MLP and GB)

Upon further evaluation via 20-fold validation, the results indicate that the GB outperforms the MLP in terms of 20-fold accuracy, which is 99% greater than the MLP accuracy of 88.9%. Similarly, the other parameters yield better results than those of the MLP. Therefore, in a nutshell, we can say that gradient boosting performs well compared with the MLP after performing cross-validation.

5.4 Model evaluation via the LIME and SHAP techniques

LIME and SHAP play important roles in model evaluation. SHAP and LIME help us analyse the decisions made by our machine learning models. ML models, including gradient boosting and MLP, even perform well in terms of prediction but are often referred to as “black boxes” because they do not explain their reasoning for the results they provide. With LIME, we can explain individual predictions by determining what impact certain features had on a certain outcome, e.g., the number of mutual friends, profile picture, etc. SHAP goes a step further by providing explanations for the contribution each feature provides to the predictions made, be it for single instances or cumulative scenarios. This helps us verify that the model is operating as expected, offers trust in the system, and facilitates the communication of the results to other stakeholders. With respect to the LIME for both models, the results are shown in Figures 17 and 18.

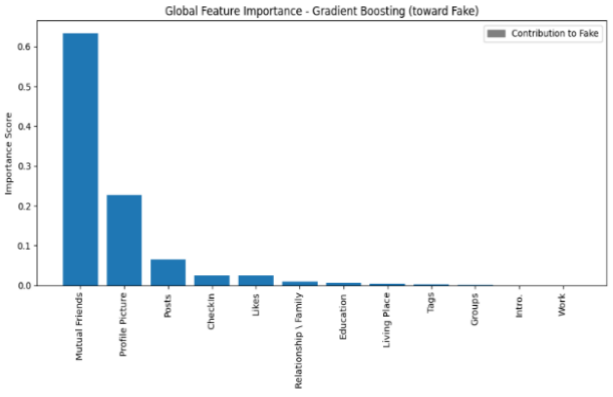

Fig. 17: Global feature importance- gradient boosting

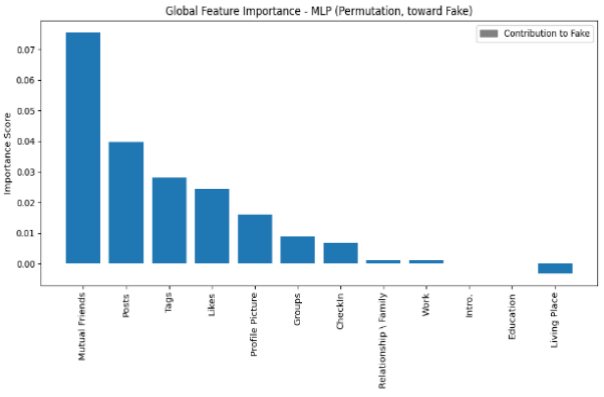

Fig. 18: Global feature importance- multilayer performance

While applying the LIME model to check which model is best fit for fake profile detection, the results below (Figures 19 and 20) show that gradient boosting covers more parameters than does MLP because the gradient boosting model concentrates on powerful features such as mutual friends and profile pictures. Unlike the MLP model, which uses many different features, such as posts, tags, likes and others, gradient boosting is considered the best model for fake profile classification.

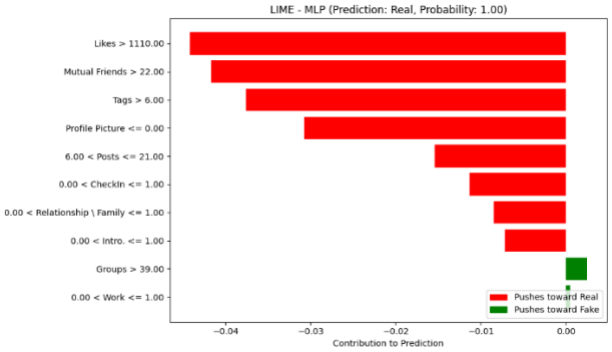

Fig. 19 Local feature importance multilayer perceptron.

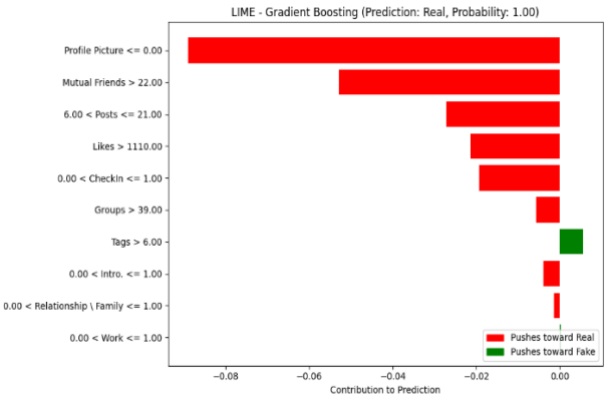

Fig. 20 Local feature importance gradient boosting.

LIME demonstrates that the gradient boosting model has a more focused and concise decision than the MLP model does. For the gradient boosting model, the most significant features driving the prediction toward a real profile are having a profile picture (profile picture ≤ 0.00) and having more than 22 common friends, with significant contributions of approximately –0.08 and –0.06, respectively. These individual features alone are strong reasons for the model's decision. In contrast, the MLP model distributes its attention across more features with diminishing individual contributions. It is dependent on features such as Likes > 1110, Mutual Friends > 22, and Tags > 6, each having a similar –0.04 contribution. Although the models predict the same outcome, the gradient boosting model employs fewer but more significant features and therefore makes more understandable decisions. This establishes that gradient boosting is not only more accurate but also more understandable and trustworthy for identifying fake profiles.

The SHAP values (Figures 21 and 22) display how each feature affects the prediction made by the model in determining whether the profiles are fake. The most significant features are profile pictures, mutual friends, and posts. Upon overall summary, those with no profile picture (value= 0) and with fewer common friends are likely to be classified as fakes. These two features are the most impactful in predictions, with SHAP values ranging from as high as +6 or as low as –4 in the case of mutual friends and strong SHAP values in the case of the profile picture. Examining the individual breakdown in one profile, the model used a foundation prediction of –5.987, and after considering the characteristics, the overall score turned to –9.39, driving the prediction more in the "fake" direction.

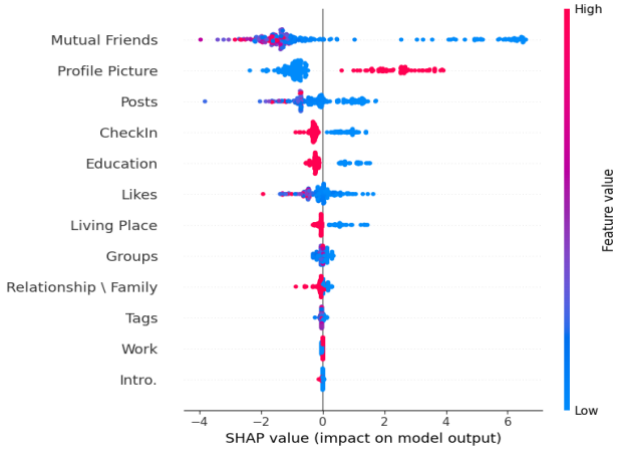

Fig. 21 SHAP Evaluation Using GB with Global Features

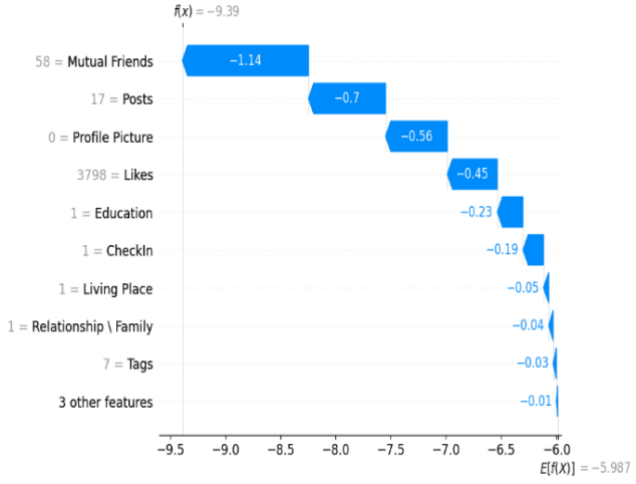

Fig. 22: SHAP Evaluation of GB with Local Feature

6. LIMITATIONS

The proposed model shows promising accuracy on a curated Facebook dataset, but several constraints limit the broader applicability of our findings. The 1244 manually verified profiles collected within a narrow time window do not represent

the complete range of adversarial behavior and language variations or the regional conventions that exist in production environments. The feature palette uses a single snapshot of profile metadata to analyse data; however, it fails to detect signals that graph-based or temporal models can identify through their exploitation of friendship-network topology and bursty posting patterns. The fake-account prevalence in real platform telemetry is much lower than the ratio used during training, which creates challenges in highly imbalanced settings. The evaluation process was restricted to supervised learning under predefined assumptions, as the test profiles had the same attribute schema as the training data, but the model was not tested against adversarial feature manipulation or novel attack vectors (e.g., AI-generated bios) that could appear after deployment. The study focuses exclusively on Facebook and does not assess the applicability of findings to other social media platforms, as differences in attribute semantics may limit portability to other social ecosystems. The identified limitations require future research to develop large-scale multimodal detection systems that can resist adversarial attacks.

7. CONCLUSION

This work advances the application of supervised learning in the domain of fake profile detection by empirically evaluating and comparing two well-established machine learning models, gradient boosting and multilayer perceptron, on a real-world social media dataset. The study underscores the superior performance and stability of gradient boosting, particularly in scenarios involving unprocessed and imbalanced data, highlighting its practical suitability for deployment in dynamic environments such as metaverse platforms. The key contribution of this research lies in its integration of explainable AI (XAI) methods, namely, LIME and SHAP, which provide model-agnostic interpretability and actionable insights into feature importance. This aligns with the growing emphasis in the AI community on transparency and accountability, especially in high-stakes applications were model decisions impact trust and security. These findings support the development of AI systems that are not only accurate but also interpretable and robust qualities essential for combating adversarial behaviors in complex social ecosystems. Future directions include extending this work with graph-based and temporal models, exploring adversarial robustness, and generalizing across multiple platforms to build more resilient and trustworthy AI solutions in the broader context of digital identity verification.

Conflicts of interest

The authors declare no conflicts of interest.

Funding

This research work was funded by Umm Al-Qura University, Saudi Arabia, under grant number 25UQU4400257GSSR13.

Acknowledgement

The authors extend their appreciation to Umm Al-Qura University, Saudi Arabia, for funding this research work through grant number 25UQU4400257GSSR13.

References

[1] H. Ning et al., "A Survey on Metaverse: the State-of-the-art, Technologies, Applications, and Challenges," arXiv:2111.096732111.09673v1, Nov. 2021. [Online]. Available: https://arxiv.org/abs/2111.09673v1

[2] P. 'Asher' Rospigliosi, "Metaverse or Simulacra? Roblox, Minecraft, Meta and the turn to virtual reality for education, socialisation and work," Interactive Learning Environments, vol. 30, no. 1, pp. 1–3, 2022. [Online]. Available: https://doi.org/10.1080/10494820.2022.2022899

[3] E. Howcroft, "Virtual real estate plot sells for record $2.4 million," Reuters, Nov. 23, 2021. [Online]. Available: https://www.reuters.com/markets/currencies/virtual-real-estate-plot-sells-record-24-million-2021-11-23/

[4] S. Mansfield-Devine, "Hacking democracy: abusing the Internet for political gain," Network Security, vol. 2018, no. 10, pp. 15–19, 2018. [Online]. Available: https://doi.org/10.1016/S1353-4858(18)30102-8

[5] W. M. S. Yafooz, A.-H. M. Emara, and M. Lahby, "Detecting Fake News on COVID-19 Vaccine from YouTube Videos Using Advanced Machine Learning Approaches, " in Combating Fake News with Computational Intelligence Techniques, M. Lahby, A. S. K. Pathan, Y. Maleh, and W. M. S. Yafooz, Eds. Cham, Springer, 2021, vol. 1001, Studies in Computational Intelligence, pp. 421–435. [Online]. Available: https://doi.org/10.1007/978-3-030-90087-8_21

[6] K. J. Brakas and M. Alanezi, "Measuring the Extent of Cyberbullying Comments in Facebook Groups for Mosul University Students," Mesopotamian Journal of CyberSecurity, vol. 5, no. 2, pp. 337–348, 2025, doi: 10.58496/MJCS/2025/021.

[7] C. Nurik, "Men Are Scum': Self-Regulation, Hate Speech, and Gender-Based Censorship on Facebook," International Journal of Communication, vol. 13, 2019. [Online]. Available: https://ijoc.org/index.php/ijoc/article/view/9608

[8] M. Conti, R. Poovendran, and M. Secchiero, "FakeBook: Detecting fake profiles in on-line social networks," in Proc. 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Istanbul, Turkey, 2012, pp. 1071–1078, doi: 10.1109/ASONAM.2012.185.
[9] R. V. Kotawadekar, A. S. Kamble, and S. A. Surve, "Automatic detection of fake profiles in online social networks," International Journal of Computer Sciences and Engineering, vol. 7, no. 7, pp. 40–45, 2019. [Online]. Available: https://doi.org/10.26438/ijcse/v7i7.4045
[10] A. Gupta and R. Kaushal, "Towards detecting fake user accounts in facebook," in Proc. 2017 ISEA Asia Security and Privacy (ISEASP), Surat, India, 2017, pp. 1–6, doi: 10.1109/ISEASP.2017.7976996.
[11] X. Wang, C.-M. Lai, Y.-C. Lin, C.-J. Hsieh, S. F. Wu and H. Cam, "Multiple Accounts Detection on Facebook Using Semi-Supervised Learning on Graphs," in Proc. MILCOM 2018 - 2018 IEEE Military Communications Conf. (MILCOM), Los Angeles, CA, USA, 2018, pp. 1–9, doi: 10.1109/MILCOM.2018.8599718.
[12] M. Albayati and A. Altamimi, "MDFP: A Machine Learning Model for Detecting Fake Facebook Profiles Using Supervised and Unsupervised Mining Techniques," International Journal of Simulation: Systems, Science & Technology, vol. 20, no. 1, 2020. [Online]. Available: https://doi.org/10.5013/IJSSST.a.20.01.11
[13] H. A. Al-Tameemi et al., "A Systematic Review of Metaverse Cybersecurity: Frameworks, Challenges, and Strategic Approaches in a Quantum-Driven Era," Mesopotamian Journal of CyberSecurity, vol. 5, no. 2, pp. 770–803, 2025, doi: 10.58496/MJCS/2025/045.
[14] G. Stringhini, C. Kruegel, and G. Vigna, "Detecting spammers on social networks," in Proc. 26th Annual Computer Security Applications Conf. (ACSAC '10). New York, NY, USA: ACM, 2010, pp. 1–9. [Online]. Available: https://doi.org/10.1145/1920261.1920263
[15] K. Thomas, D. McCoy, C. Grier, A. Kolcz, and V. Paxson, "Trafficking fraudulent accounts: The role of the underground market in twitter spam and abuse," in Proceedings of the 22nd USENIX Security Symposium, USENIX Association, Washington, DC, USA, 2013, pp. 195–210. [Online]. Available: https://www.usenix.org/conference/usenixsecurity13/technical-sessions/paper/thomas
[16] M. Jiang, P. Cui, A. Beutel, C. Faloutsos, and S. Yang, "Detecting suspicious following behavior in multimillion-node social networks," in Proc. 23rd International Conference on World Wide Web (WWW '14 Companion), New York, NY, USA:ACM, 2014, pp. 305–306. [Online]. Available: https://doi.org/10.1145/2567948.2577306
[17] M. Jiang, P. Cui, A. Beutel, C. Faloutsos, and S. Yang, "Catching synchronized behaviors in large networks: A graph mining approach," ACM Transactions on Knowledge Discovery from Data, vol. 10, no. 4, 2016. [Online]. Available: http://dx.doi.org/10.1145/2746403
[18] C. Xiao, D. M. Freeman, and T. Hwa, "Detecting clusters of fake accounts in online social networks, " in Proc. 8th ACM Workshop on Artificial Intelligence and Security (AISec '15), New York, NY, USA:ACM, 2015, pp. 91–102. [Online]. Available: https://doi.org/10.1145/2808769.2808779
[19] A. Elazab, A. M. Idrees, M. A. Mahmoud, and H. Hefny, "Fake accounts detection in twitter based on minimum weighted feature," in Proc. 18th International Conference on Document Analysis and Recognition (ICDAR), Johannesburg, South Africa, 2016. [Online]. Available:http://www.fayoum.edu.eg/English/ComputersInformation/InformationSystems/pdf/DrNesreen2.pdf
[20] Z. Yang, C. Wilson, X. Wang, T. Gao, B. Y. Zhao, and Y. Dai, "Uncovering social network sybils in the wild," ACM Transactions on Knowledge Discovery from Data, vol. 8, no. 1, pp. 1–29, 2014. [Online]. Available: https://doi.org/10.1145/2556609
[21] S. Adikari and K. Dutta, "Identifying fake profiles in LinkedIn," arXiv:2006.0138, 2020. [Online]. Available: https://arxiv.org/abs/2006.01381
[22] G. Kontaxis, I. Polakis, S. Ioannidis, and E. P. Markatos, "Detecting social network profile cloning," in Proc. *International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, Seattle, WA, USA, 2011, pp. 295–300, doi: 10.1109/PERCOMW.2011.5766886
[23] A. Shah, S. Varshney, and M. Mehrotra, "Detection of fake profiles on online social network platforms: Performance evaluation of artificial intelligence techniques," SN Computer Science, vol. 5, 2024. [Online]. Available: https://doi.org/10.1007/s42979-024-02839-9
[24] H. Ali, I. Malik, S. Mahmood, F. Akif and J. Amin, "Sybil Detection in Online Social Networks," in Proc. 2022 17th International Conference on Emerging Technologies (ICET), Swabi, Pakistan, 2022, pp. 125–129, doi: 10.1109/ICET56601.2022.10004683
[25] M. Mahapatra, N. Gupta, R. Kushwaha, and G. Singal, "Data breach in social networks using machine learning,"Advanced Computing, Garg, D., Jagannathan, S., Gupta, A., Garg, L., Gupta, S. Eds. Cham: Springer, 2021, vol. 1528, Communications in Computer and Information Science, pp. 660–670. [Online]. Available: https://doi.org/10.1007/978-3-030-95502-1_50

[26] S. R. Sahoo and B. B. Gupta, "Fake profile detection in multimedia big data on online social networks," International Journal of Information and Computer Security, vol. 12, no. 2-3, pp. 303–331, 2020. [Online]. Available: https://doi.org/10.1504/ijics.2020.105181
[27] S. Shehnepoor, R. Togneri, W. Liu, and M. Bennamoun, "Social Fraud Detection Review: Methods, challenges and analysis," arXiv:2111.05645, 2021. [Online]. Available: https://arxiv.org/abs/2111.05645
[28] H. Taherdoost, "Enhancing Social Media Platforms with Machine Learning Algorithms and Neural Networks," Algorithms, vol. 16, no. 6, Art. no. 271, 2023. [Online]. Available: https://doi.org/10.3390/a16060271
[29] M. Chakraborty, S. Das, and R. Mamidi, "Detection of fake users in SMPs using NLP and graph embeddings," arXiv:2104.13094, 2021. [Online]. Available: http://arxiv.org/pdf/2104.13094.pdf
[30] J. Ezarfelix, N. Jeffrey, and N. Sari, "A Systematic Literature Review: Instagram Fake Account Detection Based on Machine Learning", Engineering, MAthematics and Computer Science Journal (EMACS), vol. 4, no. 1, pp. 25–31, 2022. [Online]. Available: https://doi.org/10.21512/emacsjournal.v4i1.8076
[31] K. Shu, "Combating Disinformation on Social Media and Its Challenges: A Computational Perspective", in Proc. AAAI Conference on Artificial Intelligence, vol. 37, no. 13, p. 15454, 2024. [Online]. Available: https://doi.org/10.1609/aaai.v37i13.26821
[32] Y. Liu, Z. Boukouvalas, and N. Japkowicz, "A semi-supervised framework for misinformation detection," in Discovery Science. DS 2021, Soares, C., Torgo, L. Eds. Cham: Springer, 2021, vol. 12986, Lecture Notes in Computer Science, pp. 57–66. [Online]. Available: https://doi.org/10.1007/978-3-030-88942-5_5
[33] M. Fire, D. Kagan, A. Elyashar, and Y. Elovici, "Friend or foe? Fake profile identification in online social networks," Social Network Analysis and Mining, vol. 4, Art. no. 194, 2014. [Online]. Available: https://doi.org/10.1007/s13278-014-0194-4
[34] A. Nazir, S. Raza, C.-N. Chuah, and B. Schipper, "Ghostbusting Facebook: Detecting and Characterizing Phantom Profiles in Online Social Gaming Applications," in Proc. 3rd Workshop *on Online Social Networks (WOSN 2010)*, Boston, MA, USA, 2010. [Online]. Available: https://www.usenix.org/conference/wosn-2010/ghostbusting-facebook-detecting-and-characterizing-phantom-profiles-online
[35] M. McCord and M. Chuah, "Spam detection on Twitter using traditional classifiers," in Autonomic and Trusted Computing (ATC 2011), Calero, J.M.A., Yang, L.T., Mármol, F.G., García Villalba, L.J., Li, A.X., Wang, Y., Eds. Berlin: Springer, 2011, vol 6906, Lecture Notes in Computer Science. [Online]. Available: https://doi.org/10.1007/978-3-642-23496-5_13
[36] J. H. Hong, E. K. Yun, and S. B. Cho, "A Review of Performance Evaluation for Biometrics Systems," International Journal of Image and Graphics, vol. 5, no. 3, pp. 501–536, 2005. [Online]. Available: https://doi.org/10.1142/S0219467805001872