RESEARCH ARTICLE

Machine learning for detecting fake accounts and genetic algorithm-based feature selection

Amine Sallah[1] [iD], El Arbi Abdellaoui Alaoui[2], Stéphane C.K. Tekouabou[3,4] [iD] and Said Agoujil[5]

[1]Department of Computer Science, Faculty of Sciences and Techniques, Moulay Ismail University, Meknes, Morocco
[2]Department of Sciences, Ecole Normale Supérieure, Moulay Ismail University, Meknes, Morocco
[3]Research Laboratory in Computer Science and Educational Technologies (LITE), University of Yaoundé 1, Yaoundé, Cameroon
[4]Department of Computer Science and Educational Technologies (DITE), Higher Teacher Training College (HTTC), University of Yaoundé 1, Yaoundé, Cameroon
[5]Department of Sciences, École Nationale de Commerce et de Gestion, Moulay Ismail University, Meknes, Morocco
Corresponding author: Amine Sallah; Email: aminefste@gmail.com

Abstract

People rely extensively on online social networks (OSNs) in Africa, which aroused cyber attackers' attention for various nefarious actions. This global trend has not spared African online communities, where the proliferation of OSNs has provided new opportunities and challenges. In Africa, as in many other regions, a burgeoning black-market industry has emerged, specializing in the creation and sale of fake accounts to serve various purposes, both malicious and deceptive. This paper aims to build a set of machine-learning models through feature selection algorithms to predict the fake account, increase performance, and reduce costs. The suggested approach is based on input data made up of features that describe the profiles being investigated. Our findings offer a thorough comparison of various algorithms. Furthermore, compared to machine learning without feature selection and Boruta, machine learning employing the suggested genetic algorithm-based feature selection offers a clear runtime advantage. The final prediction model achieves AUC values between 90% and 99.6%. The findings showed that the model based on the features chosen by the GA algorithm provides a reasonable prediction quality with a small number of input variables, less than 31% of the entire feature space, and therefore permits the accurate separation of fake from real users. Our results demonstrate exceptional predictive accuracy with a significant reduction in input variables using the genetic algorithm, reaffirming the effectiveness of our approach.

Policy Significance Statement

Machine-learning algorithms coupled with genetic algorithm-based feature selection offer a powerful approach for detecting fake accounts in Online Social Networks platforms. This research demonstrates that by leveraging advanced machine-learning techniques and employing genetic algorithms for feature selection, it is possible to achieve highly accurate and efficient identification of fake accounts. Furthermore, the integration of genetic algorithm-based feature selection optimizes the performance of the detection system by identifying the most informative features. This improves the efficiency and effectiveness of fake account detection and reduces computational complexity.

1. Introduction

As technology has advanced, Facebook, LinkedIn, Twitter, Instagram, and other social media networks have been overgrown during the previous century. In addition, smartphones, tablets,

computers, and other contemporary technologies make wireless communication more effective and convenient.

In January 2023, Northern and Southern African nations claimed the highest proportion of social media users on the continent. According to Statista (Galal 2023), statistics report in Northern Africa, approximately 49% of the population engaged with social media, whereas in Southern Africa, this metric reached 41.3%. Conversely, Central Africa exhibited the lowest social media usage rate in Africa, with only around 7% of its population participating, representing both the lowest figure across the continent and the smallest regional presence globally.

A fake profile is a cloned or malicious account. When a user impersonates another user's account, we refer to it as a "duplicate." Hiding a malicious account's true identity for malicious activity has developed considerably in recent years. Making use of fraudulent accounts damages one's reputation, causing unneeded perplexity through unexpected received messages (Sheikhi, 2020).

First, it should be highlighted that Facebook has developed an immune system called (Carmi, 2020) to combat issues caused by fake accounts by building classifiers. Every read and written operation is subject to real-time inspections and classifications by this immune system. In situations involving Advanced Persistent Threats, fake identities on social media are frequently used to transmit malwares or suspicious links. Furthermore, they are employed in various nefarious activities like sending spams and spam emails, and in certain applications, to promote and inflate the number of users (Joshi et al., 2020).

In this research, we have attempted to detect active fake accounts on Facebook and Instagram. Furthermore, we have applied feature selection to decrease the number of variables and reduce computation time.

Typically, machine-learning classifiers employed for the prediction of fake accounts can be classified into the following categories (Singhal et al., 2018):

1. Bagging, which frequently considers homogenous weak learners, who learn from one another both in parallel and separately, and then aggregates them using some deterministic averaging procedure.
2. Boosting, which often includes homogenous weak learners, learns them sequentially in a highly adaptative manner (each base model is dependent on the preceding ones) and combines them using a deterministic technique.

The aforementioned machine-learning models are based on input data that includes attributes characterizing the objects under investigation. In the machine-learning language, a feature is a variable, parameter or property. Therefore, feature selection significantly impacts model performance in data-driven models.

In recent years, datasets have commonly included numerous attributes, encompassing both pertinent and superfluous information. However, the existence of redundant and unimportant attributes not only expands the feature space but also hinders the effectiveness of machine-learning techniques (Cai et al., 2018; Venkatesh and Anuradha, 2019). Consequently, feature selection emerges as an effective approach to reduce dimensionality issues. This method, performed as a pre-processing step, enhances data interpretability by carefully selecting a subset of significant attributes. The primary purpose of feature selection is to discover the optimal combination of relevant features that may significantly improve classification performance while reducing classification model complexity.

For statistical pattern recognition, data mining, and machine learning, feature selection has been a promising field of study and development since the 1970s. As a result, many attempts have been made to assess the feature selection techniques, which can be divided into four groups based on the evaluation process, namely, filters, wrappers, hybrids, and embedded (Rostami et al., 2021). A filter technique conducts feature selection independently of any learning algorithm (for example, a totally separate preprocessing). The feature set filter strategy requires statistical analysis, which can only be used to handle the feature selection problem without using a learning model. On the other hand, the wrapper technique uses a predefined learning process to determine the quality of the chosen subsets. Furthermore, Wrappers may produce superior results but are more expensive to run and can disintegrate with too many

features. The hybrid strategy uses a combination of both filter and wrapper techniques. Finally, the embedded approaches benefit from selecting features in the learning process and are equivalent to a specific learning model (Liu et al., 2018; Moslehi and Haeri, 2020).

The BORUTA algorithm (Kursa and Rudnicki, 2020) was used to select the specified characteristics. Given that reducing the number of input variables is an essential aspect of building a predictive model, it was also determined that Boruta should explore whether, during the construction of a predictive model, it is necessary to select all the features identified as relevant by the BORUTA algorithm, or if it is possible to reduce the number of features and assess how such reduction impacts the accuracy of fake account prediction. Evaluation criteria such as AUC, Accuracy, Precision, and MCC were used to evaluate the quality of the built prediction model (Rácz et al., 2019).

Another helpful approach is the genetic algorithm (GA), which provides optimized output through distinct feature analyses for the input data. J.H. Holland suggested GA in 1992. Chromosome representation, fitness, selection, and biological-inspired operators are the fundamental parts of GA (Katoch et al., 2021). This strategy is more efficient and effective in terms of efficiency and performance. The genetic algorithm (GA) has been extensively employed in feature selection problems as a basic optimization method. However, one of the weaknesses of GA is hyperparameter tuning, high computing cost, and randomness of the selection procedure (Allam and Nandhini, 2018).

The remainder of this paper is organized as follows: In Section 2, we discuss previous works in the area of detecting fake accounts in online social networks and feature selection methods. Section 3 discusses our fake detection method and provides experimental findings. The study concludes with a summary of the findings and some conclusions in Section 4.

2. Related Work

In recent years, feature selection has emerged as a critical research component in machine learning; the focus areas encompass image retrieval, text mining, intrusion detection, and other fields. As a result, various methods for feature selection have been developed and used in the literature (Deng et al., 2019; Adewole et al., 2021; Anand et al., 2021). For example, such as particle swarm optimization (PSO) (Shami et al., 2022), ant colony optimization (ACO) (Ma et al., 2021), and GA (Ghatasheh et al., 2022). GA has long been recognized as a very effective and practical method for feature selection, as described in Fraser et al. (1970), Deepa (2008). This is due to its capacity to change the functional configuration for better performance outcomes. Online social networks have attracted various types of undesirable activities, and the academic community has already provided a variety of remedies to the issue, which are focused on profile-based (Kaubiyal and Jain, 2019), emotion-based (Wani et al., 2019), graph-based (Mohammadrezaei et al., 2018; Wang et al., 2019) and behavioral-based (Bhattasali and Saeed, 2021) features, including:

1. Profile-based: account age, gender, relationship status, education details, country, followers Count and friends Count.
2. Graph-based: betweenness centrality, in/out-degree, Friendship evolution, OSN graph structure evolution, clustering coefficient, connection strength, etc.
3. Content-based: URLs in the post, message, similarity, message length, hashtags, number of tags, punctuation count, number of capital letter words, sentence length, and so forth.
4. Behavioral-based: Posts sent in a particular time interval, time in days.

Gupta and Kaushal (2017) attempted to identify fraudulent accounts on Facebook based on user profile activity and interactions with other users. These actions were defined by an extensive feature set that included Facebook users'like, comment, share, tag, and app use patterns. They then used their dataset to apply the most widely used supervised machine-learning classification algorithms.

The suggested strategy was based on establishing the effective features for the identification procedure in Munga and Mohandas (2022). Then, the obtained features were filtered using entropy and information gain. As a result, the suggested technique contains just eight effective features for fraudulent account identification out of 25 attributes, with five decided features based on information gain.

Sahoo and Gupta (2020) presented a hybrid solution that uses several machine-learning algorithms to distinguish between spammer and non-spammer contents and accounts. Initially, the genetic algorithm analyzes the numerous variables and picks the most relevant features that impact the behavior of user accounts, which are then used for training classifiers. As a result, their system was successful in effectively differentiating spammer and non-spammer content. Finally, a comparison with certain current state-of-the-art methodologies is performed to demonstrate the effectiveness of the suggested framework. According to the experimental results, their strategy produces a high detection rate of 99.6%, which is superior to other state-of-the-art approaches.

Akyon and Esat Kalfaoglu (2019) gathered datasets to identify fake and automated accounts, introducing derived characteristics for classifying fake and automated accounts. Additionally, they implemented a cost-sensitive feature reduction approach using genetic algorithms to optimize the selection of features for automated account classification. Furthermore, they used the SMOTE-NC algorithm to rectify the unevenness in the fake account dataset and evaluate multiple pattern recognition algorithms on the obtained datasets. As a result, SVM obtained an F1 score of 86% for automated account identification, and the neural network achieved an F1 score of 95%.

3. Proposed Methodology

Our study explores the application of genetic algorithms for feature selection in the context of fake account detection. Our objective is to identify the most relevant features to be considered in the detection process. The workflow for this research is shown in Figure 1. The experiment comprises several steps. Firstly, we initiated our work by collecting and processing the dataset. Secondly, we employed the GA and Boruta methods for selecting key features. Next, we compared seven classifier methods using the selected features to determine the optimal classification approach for predicting unknown user accounts.

3.1. Dataset and features

Many datasets are openly accessible on the Internet, and we can adopt them for Instagram and Facebook fake account classification. Figure 2 shows how each set is balanced for the target class.
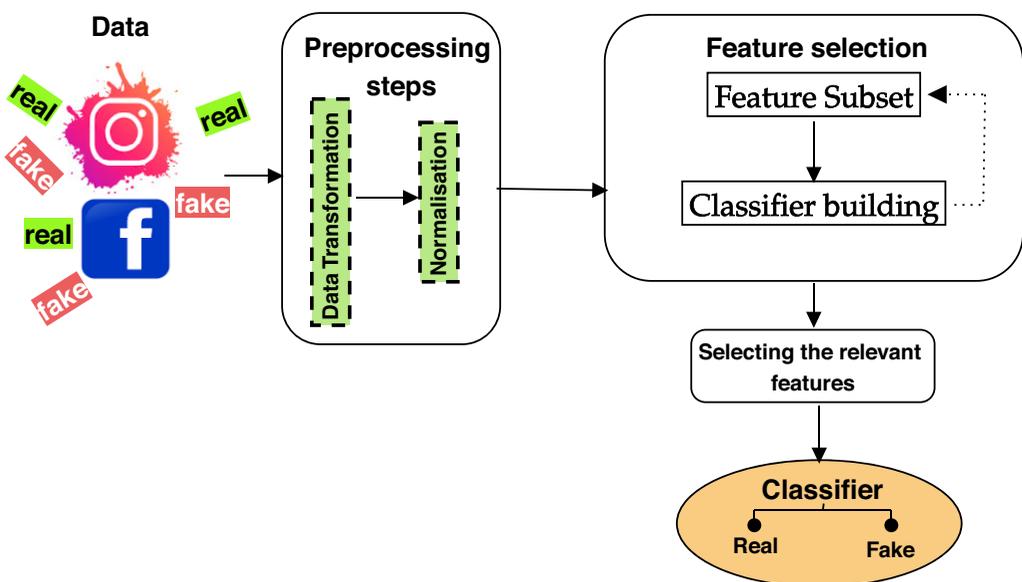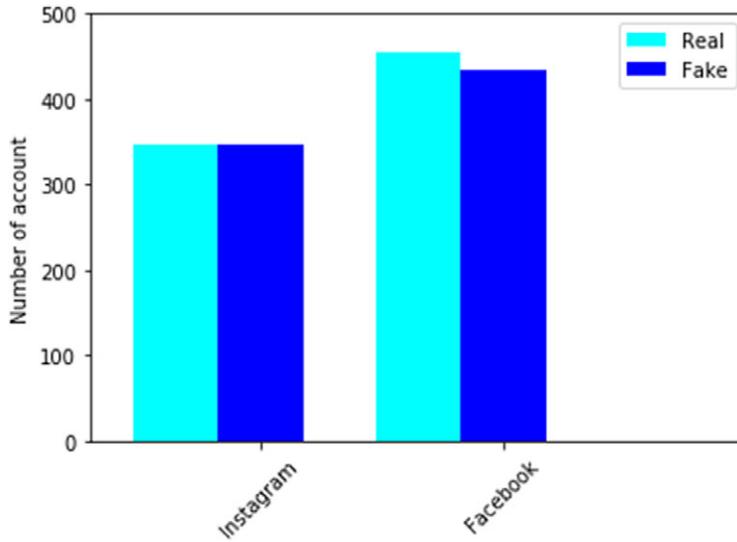

Figure 1. *Proposed system detection.*

Figure 2. *Count plot of the number of instances of the partitions of the Facebook and Instagram dataset for each class.*

3.1.1. Instagram

A CSV file with 699 rows and 12 columns, including the target variable, was downloaded as the dataset. The source of this Dataset is Bakhshandeh (2019). The target variable accepts the values genuine and fake, allowing us to determine if the account is fraudulent. A collection of 11 characteristics were given as stated, according to Table 1.

Before constructing any machine-learning model for a tabular dataset, assessing the association between the independent and target variables is common practice. This assessment typically involves quantifying the correlation between the two variables. For this purpose, we exploit a Pearson correlation, a number between -1 and 1 that indicates the extent to which two variables are linearly related. The heat map data in Figure 3 reveal no high association between the variables.

3.1.2. Facebook

The second dataset comprises profiles of two statutes: real and fake. We used the dataset "Fake and Real Accounts Fakebook" (Albayati and Altamimi, 2019) for our research. The collection includes data from

Table 1. *Features that characterize each user profile on Instagram*

| Feature name	Description	Feature type
Profil pic	User has a profile picture or not	Categorical
Nums/length username	The ratio of numeric characters to the length of the username	Numeric
Fullname words	Number of words in the full name	Numeric
Nums/length fullname	Percentage of numeric characters in the full name.	Numeric
Name==username	Are the username and full name the same?	Categorical
Description length	Biography length in characters	Numeric
External URL	Has an external URL or not?	Categorical
Private	Private or not	Categorical
#Posts	Number of publications	Numeric
#Followers	Number of subscribers	Numeric
#Follows	Number of following	Numeric
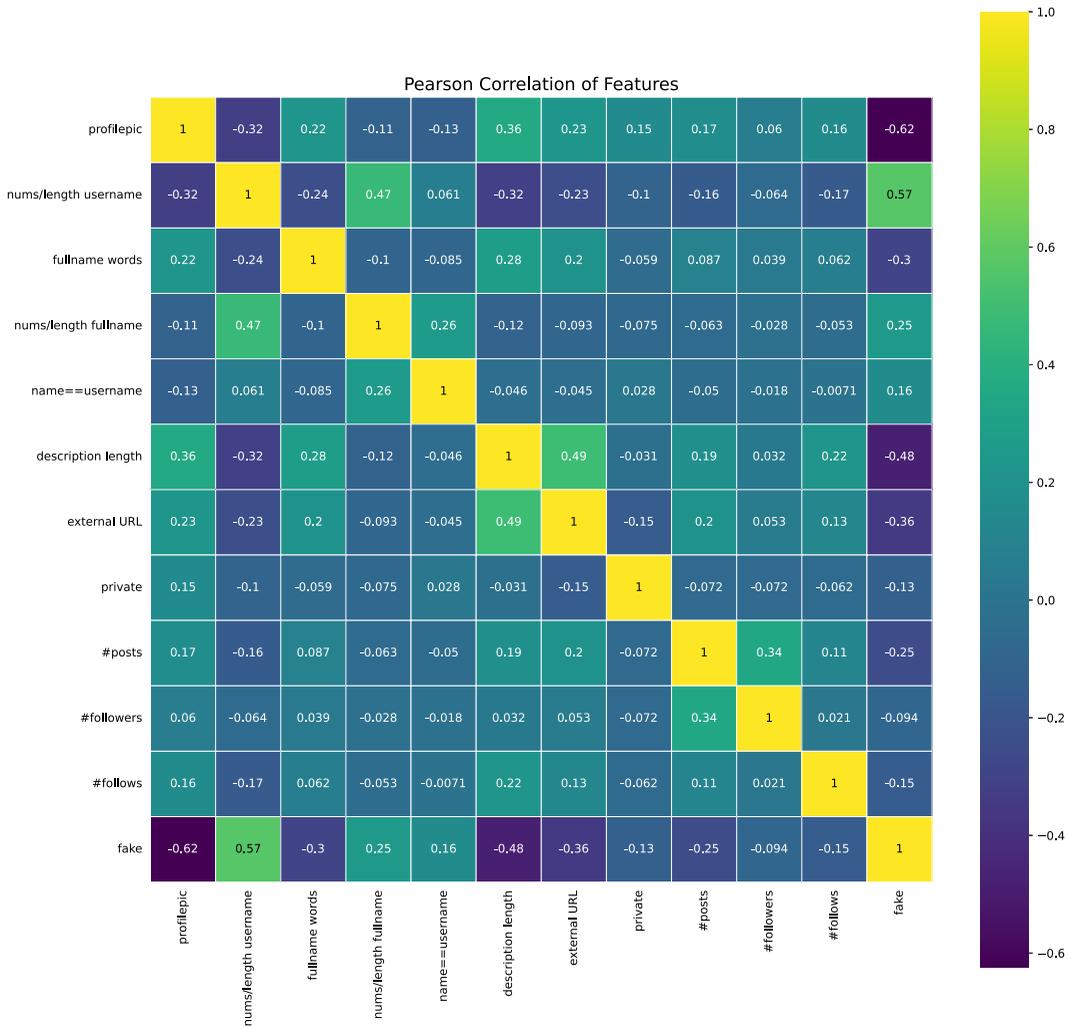
Figure 3. *Heat map data and the correlation between different variables in Instagram.*

889 public Facebook accounts. It consists of 22 predictor variables and one outcome variable (Status), indicating whether the account is legitimate or fake. In Figure 4, we provide a quick overview of all the predictor variables.

We initially analyzed the Facebook dataset for multicollinearity concerns among our predictor variables. This is interpreted as follows: a correlation value of 0.7 between two variables indicates that the two have a significant and positive association (Nettleton, 2014). Consider Figure 5, which depicts the correlation between 23 variables.

Figure 4. *Screenshot of Facebook dataset.*
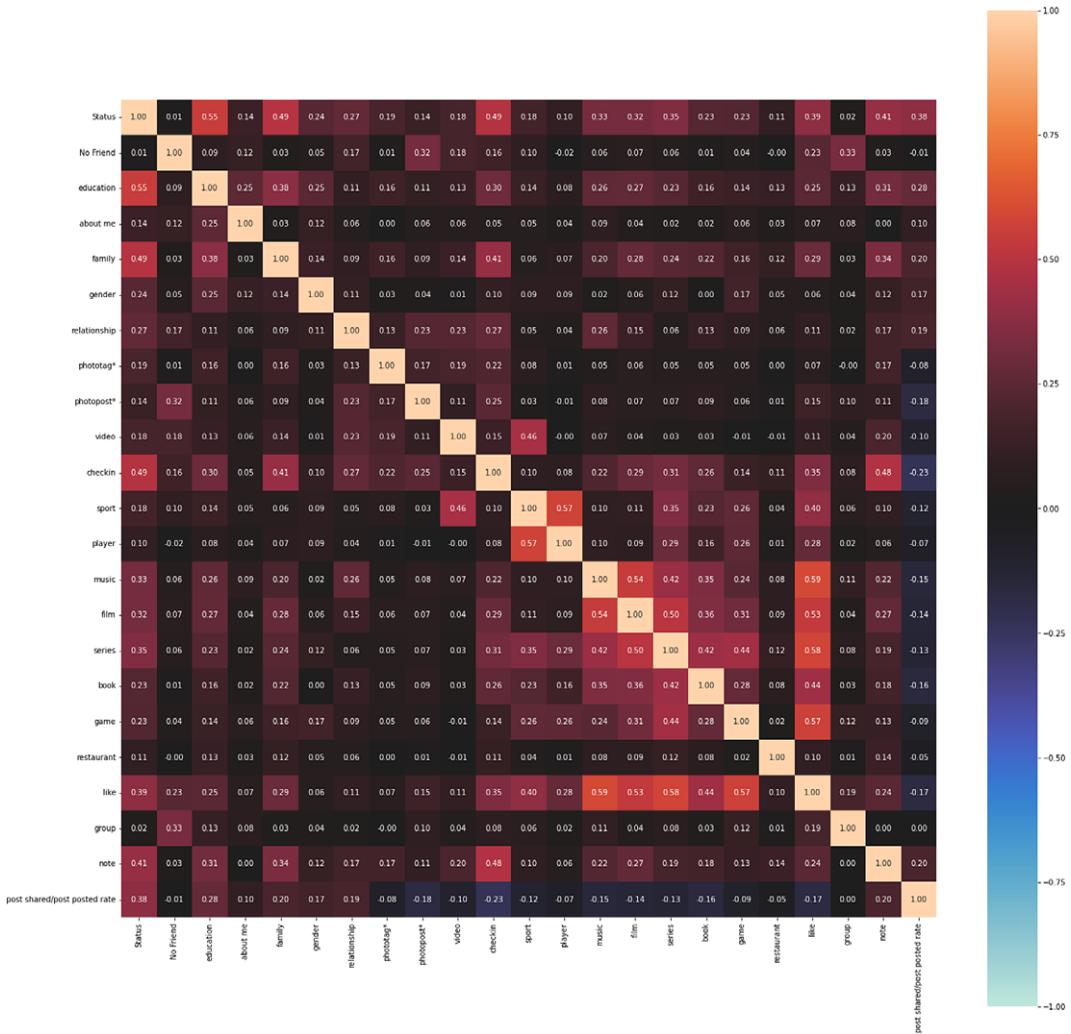
Figure 5. *Heat map data and the correlation between different variables in Facebook.*

3.2. Feature selection

It is rare for all the dataset's variables to be employed in the development of a machine-learning model. Unnecessary variables impact a model's capacity to generalize and might lower a classifier's overall accuracy. Additionally, increasing the number of variables in a model makes it more complicated overall (Gazeloglu, 2020).

3.2.1. Genetic algorithm

GA (Jennings et al., 2019) is an evolutionary computation technique that consists of methods for solving multi-objective optimization problems, GA is used to find the optimal combination of feature subsets and parameters.

The terms `population` and `generation` refer to groups of people, `generation` is an iteration of the optimization (evolutionary) process, and `multi-objective fitness function` $m \in \mathbb{N}$ refers to a collection of real functions $g_i : S$ for $i \in \{1, \ldots, m\}$, where each function $g_i :$ assesses a different quality of a given individual. A person belongs to the set S of all prospective individuals who may be

formed for a certain problem. A standard generation of an evolutionary process based on GA consists of the phases listed below:

1. Fitness evaluation: The fitness function processes the execution using specific classification algorithms by choosing various features. Here, we examine and validate each feature related to the Facebook and Instagram accounts using a classifier. Then, the classifier's performance and other features in various chromo are used to estimate the fitness value;
2. SELECTION: based on the fitness value threshold cut-off, multiple fitness chromosomes are chosen. Chromosomes with greater probability are selected for the following stage of procedures;
3. CROSSOVER: it includes fusing and combining the data structures of two parents to create children;
4. MUTATION: Each bit in a chromosome with binary values typically has a predetermined probability of flipping.

The standard procedure of a GA is shown in Figure 6.

3.2.2. *Boruta*

Boruta is a method for selecting and prioritizing features that relies on the Random Forest algorithm. The advantages of Boruta include determining the relevance of the variable and statistically selecting important variables.

3.2.2.1. How does Boruta algorithm work?. To begin, Boruta algorithm initiates the process by augmenting the provided dataset with randomly generated duplicates of all the features, referred to as Shadow Features.

Subsequently, using this expanded dataset (comprising both original attributes and shadow attributes), Boruta algorithm trains a random forest classifier and evaluates the significance of each feature by employing a feature importance metric like Mean Decrease Accuracy.

Boruta Algorithm scrutinizes the relative importance of each actual feature at each iteration, systematically discarding features that are deemed nonessential, all while retaining the best-performing shadow features.
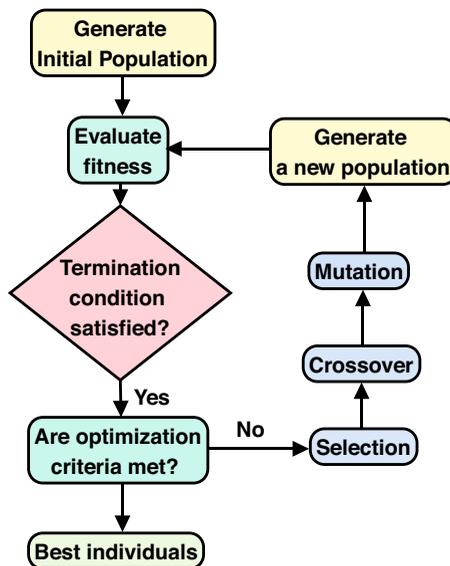

Figure 6. *Standard procedure of genetic algorithm.*

Finally, Boruta Algorithm concludes when all features have been validated or rejected or when it nears a predetermined threshold associated with the random forest.

In contrast, Boruta identifies all features that exhibit varying degrees of correlation with the target variable, whether strong or weak.

3.3. Performance metrics

The effectiveness of each approach is evaluated across multiple performance metrics, including precision, recall, F1-score, accuracy, and the Area Under the Receiver Operating Characteristic Curve (AUC/ROC). In this section, we provide concise but comprehensive explanations of each performance metric that we have used in our research.

1. *Accuracy* measures the proportion of correctly classified instances among all instances in a classification proble:

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Negative + False\ Positive}$$

2. *Precision* assesses the ability of a classifier to correctly identify positive instances. It is the ratio of true positive predictions to the total positive predictions.

$$Precision = \frac{True\ Positive}{True\ Positive + FalseP\ ositive}$$

3. *Recall* measures the ability of a classifier to identify all relevant instances. It is the ratio of true positive predictions to the total actual positive instances.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

4. *F1-Score* evaluates the trade-off between recall and precision and can be computed as follows:

$$F1 - Score = 2 \times \frac{(precision \times recall)}{(precision + recall)}$$

5. *AUC* (area under the ROC curve):

   Machine-learning classification techniques that yield a score within the range $[0, 1]$, as opposed to a binary class assignment $\{0, 1\}$, can benefit from optimization through the selection of a threshold value that may differ from 0.5. This threshold is employed to convert the continuous score into a categorical prediction. Each potential threshold value is associated with both a true positive rate (*TPR*) and a false positive rate (*FPR*). Assessing the inherent quality of the score can be accomplished by means of an ROC (Receiver Operating Characteristic) curve. This curve illustrates, for each envisageable threshold, the *TPR* in relation to the *FPR*.

The true positive rate (TPR), which is equivalent to the recall, is defined as follows:

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

False positive rate (FPR) is defined as follows:

$$FPR = 1 - \text{Specificity} = \frac{FP}{FP + TN}$$

To establish the coordinates along the Receiver Operating Characteristic (ROC) curve, it is conceivable to conduct multiple model evaluations while altering classification thresholds; however, this approach would be deemed inefficient. A commonly employed metric for assessing performance is the Area Under the ROC Curve (AUC), which assumes values within the range of 0 to 1. A higher AUC value approximating 1 signifies the model's exceptional predictive capability.

3.4. Models construction

We used seven classification algorithms to train our models: Gradient Boosting, Extreme Gradient Boosting, catBoost, Random Forest, LightGBM, AdaBoost, and ExtraTree Classifier. First, we train our model with all of its features. Then, we use the two-feature selection approach and choose important features and train our model with the chosen features. Finally, various measures are used to examine the suggested model, including accuracy, recall, precision, F1-score, and others.

We performed a 10-fold cross-validation to evaluate our models. We began by dividing the total number of observations into 10 random subsets. Next, we trained a model using nine subsets for each iteration, then tested the fitted model on the remaining subset. We then averaged the cross-validated performance measures over the 10 iterations.

4. Results

This part presents the methodology used to evaluate the effectiveness of detecting fake Instagram and Facebook profiles. Initially, we predicted fake accounts with seven classifiers and no selection. Then, we use two feature selection methods, and after picking essential features, we reapply to the classifier described before. Here, we want to demonstrate how the feature selection approach influences the prediction outcome.

4.1. Experimental setup

In this study, the development of a machine-learning classifier and the incorporation of feature selection methods based on Genetic and Boruta algorithms were accomplished through the utilization of Python scripts.

The National Center for Scientific and Technical Research (CNRST), Rabat, Morocco, supplied computing resources for this study via HPC-MARWAN (hpc.marwan.ma). The National Center for Scientific and Technical Research provides Moroccan researchers with a remotely accessible High-Performance Computing (HPC) infrastructure. The infrastructure consists of 38 nodes, each of which has the following capacity:

1. 1672 CPU Cores (165 TFlops)
2. 4 GPUs
3. 396 TB Storage
4. TB RAM

To enhance the replicability and reproducibility of our study, we provide detailed information to allow others to replicate our experiments. We also added a summary table (see Table 2) containing the algorithms' hyperparameters. Here are specific steps to achieve this:

1. GeneticSelectionCV is a module of the scikit-learn library in Python. It is a feature selection technique that uses a genetic algorithm to optimize the feature subset for a machine-learning model while performing cross-validation.
2. BorutaPy library is a feature selection method for machine learning. The main objective of BorutaPy is to identify relevant features from a potentially large pool of features, improving the model's performance and reducing the risk of overfitting.

Table 2. *Hyperparameters used for implementing feature selection methods and classifiers*

Classifier	Parameters
GeneticSelectionCV	scoring = "accuracy", n_population = 100, crossover_proba = 0.5, mutation_proba = 0.2, n_generations = 50, crossover_independent_proba = 0.5, mutation_independent_proba = 0.04, tournament_size = 3, n_gen_no_change = 10, caching = True, n_jobs = −1
BorutaPy	n_estimators = 'auto', verbose = 2, random_state = 1
XGBClassifier	learning_rate = 0.06, n_estimators = 200, max_depth = 5, min_child_weight = 5, gamma = 0, subsample = 0.8, colsample_bytree = 0.75, objective = 'binary:logistic', scale_pos_weight = 1, reg_alpha = 0.05
RandomForestClassifier	bootstrap = True, criterion = 'gini', min_impurity_decrease = 0.0, min_samples_leaf = 1, min_samples_split = 2, n_estimators = 100,
LGBMClassifier	bagging_fraction = 1.0, bagging_freq = 1, boosting_type = 'gbdt', class_weight = None, colsample_bytree = 1.0, feature_fraction = 1.0, importance_type = 'split', learning_rate = 0.2, max_depth = −1, min_child_samples = 36, min_child_weight = 0.001, n_estimators = 210, num_leaves = 6, reg_alpha = 1, reg_lambda = 0.3, silent = 'warn'
GradientBoostingClassifier	ccp_alpha = 0.0, criterion = 'friedman_mse', learning_rate = 0.01, loss = 'deviance', max_depth = 3, min_samples_leaf = 1, min_samples_split = 2, min_weight_fraction_leaf = 0.0, n_estimators = 100, validation_fraction = 0.1
AdaBoostClassifier	algorithm = 'SAMME.R', base_estimator = None, learning_rate = 1.0, n_estimators = 50, random_state = 0
CatBoostClassifier	iterations = 90, learning_rate = 0.1, depth = 7, l2_leaf_reg = 8, model_size_reg = 0.5, rsm = 1, loss_function = 'Logloss', nan_mode = 'Min', leaf_estimation_iterations = 10, leaf_estimation_method = 'Newton', random_strength = 0.2, eval_metric = 'Logloss', boosting_type = 'Plain', bootstrap_type = 'MVS'
ExtraTreesClassifier	bootstrap = False, ccp_alpha = 0.0, class_weight = 'balanced', criterion = 'entropy', max_depth = 4, max_features = 1.0, min_impurity_decrease = 0.001, min_samples_leaf = 2, min_samples_split = 10, n_estimators = 70, oob_score = False

3. Scikit-learn is a popular machine-learning library in Python that includes ensemble methods like Random Forest, AdaBoost, and Gradient Boosting, which combine multiple models to improve performance.

4.2. Results obtained before applying feature selection method

The primary goal of using these evaluation metrics is to determine how well a machine-learning model will perform on unseen data. For balanced datasets, metrics like accuracy, precision, and recall are useful for evaluating classification models. However, if the data is unbalanced, ROC/AUC provides a more accurate evaluation of model performance. Table 3 shows the result of classifiers. This table shows that catBoost yields the highest AUC on Instagram when it uses all features. In addition, Extra Trees achieved the best AUC using the Facebook dataset. Conversely, catboost consumes more execution time among all the classifiers before applying any feature selection method.

4.3. Result obtained after applying Boruta feature selection method

The Instagram dataset with 11 attributes was selected in the initial step of developing the prediction model to uncover all the important features impacting the fake account prediction. The outcomes of the selection are presented in Table 4. As shown in the table below, BORUTA highlighted the relevance of characteristics in the dataset. In this situation, 2 of the 11 features are rejected. As indicated in Table 5, Boruta uses 63% of the Facebook features offered. Following feature selection, the predictive model is built by considering the feature selection outcomes. Table 6 displays the outcome of classifiers after applying Boruta. According to the table, the classifiers' AUCs stay consistent when compared to the Non-Selection technique, except for the Random Forest classifier, which obtained a 4% improvement in AUC when the Boruta feature selection was applied to the Instagram dataset.

4.4. Result obtained after applying genetic algorithm feature selection method

By applying a genetic algorithm to choose the best combination of features with better accuracy than the baseline, We find an acceptable solution to reaching a certain number of generations. In our example, GA will stop after 50 generations.

Table 3. *Overall results of our experiment using full feature(selection)*

	Classifier	Accuracy	AUC	Recall	Precision	F1	MCC	Time (Sec)
Instagram	Ada Boost	0.9136	0.9633	0.9156	0.9119	0.9124	0.8297	**0.117**
	Random Forest	0.942	0.942	0.971	0.918	0.944	0.887	0.569
	XGBoost	0.9299	0.9828	0.9281	0.9297	0.9283	0.8608	0.387
	Extra Trees	0.9177	0.9764	0.8900	0.9389	0.9128	0.8376	0.469
	LightGBM	**0.9443**	0.9835	**0.9406**	**0.9455**	**0.9426**	**0.8894**	0.120
	Gradient Boosting	0.9340	0.9792	0.9239	0.9415	0.9320	0.8690	0.126
	CatBoost	0.9361	**0.9841**	0.9239	0.9445	0.9337	0.8729	2.214
Facebook	Ada Boost	0.9903	0.9983	0.9874	0.9938	0.9905	0.9809	0.134
	Random Forest	**0.9952**	0.9991	**0.9936**	**0.9970**	**0.9952**	**0.9905**	0.520
	XGBoost	0.9887	0.9992	0.9905	0.9879	0.9890	0.9778	0.381
	Extra Trees	0.9887	**0.9996**	0.9905	0.9877	0.9890	0.9777	0.473
	LightGBM	0.9903	0.9994	0.9905	0.9911	0.9906	0.9811	**0.085**
	Gradient Boosting	0.9904	0.9992	0.9905	0.9908	0.9906	0.9809	0.317
	CatBoost	**0.9952**	0.9992	**0.9936**	**0.9970**	**0.9952**	**0.9905**	2.701

Table 4. *Ranking features by Boruta algorithm in Instagram*

Features	Ranking	Decision
Profilepic	1	Confirmed
Nums/length username	1	Confirmed
Fullname words	1	Confirmed
Nums/length fullname	1	Confirmed
Name==username	2	Rejected
Description length	1	Confirmed
External URL	1	Confirmed
Private	3	Rejected
Posts	1	Confirmed
Followers	1	Confirmed
Follows	1	Confirmed

Table 5. *Ranking features by Boruta algorithm in Facebook*

Features	Ranking	Decision
No friend	1	Confirmed
Phototag*	1	Confirmed
Photopost*	1	Confirmed
Video	1	Confirmed
Checkin	1	Confirmed
Sport	1	Confirmed
Player	1	Confirmed
Music	1	Confirmed
Film	1	Confirmed
Series	1	Confirmed
Book	1	Confirmed
Game	2	Rejected
Restaurant	1	Confirmed
Like	1	Confirmed
Group	1	Confirmed
Post shared/post posted rate	1	Confirmed
Education_no	1	Confirmed
Education_secondary school	7	Rejected
Education_university	1	Confirmed
About me_yes	1	Confirmed
Family_yes	1	Confirmed
Gender_male	4	Rejected
Relationship_complicate	5	Rejected
Relationship_married	6	Rejected
Relationship_vide	8	Rejected
Note_yes	3	Rejected

The GA feature selection approach selects the most relevant features. Table 7 shows that the *followers, fullname words,description length,posts and follows* are essentials for predicting fake account in Instagram.

Concerning Facebook, The main features chosen using the GA feature selection methodology are visualized in Table 8. According to the outcome of selection result, it can be concluded that the important

Table 6. *Overall results of our experiment using Boruta*

	Classifier	Accuracy	AUC	Recall	Prec.	F1-score	Time (s)
Instagram	Ada Boost	0.9074	0.9621	0.9031	0.9106	0.9052	**0.117**
	Random Forest	0.9279	0.9827	**0.9281**	0.9270	0.9265	0.522
	XGBoost	0.9258	0.9796	0.9239	0.9254	0.9243	0.385
	Extra Trees	0.9238	0.9760	0.8986	0.9431	0.9196	0.477
	LightGBM	**0.9382**	0.9821	**0.9281**	**0.9450**	**0.9359**	0.133
	Gradient Boosting	0.9259	0.9790	0.9239	0.9272	0.9244	0.118
	CatBoost	0.9300	**0.9833**	0.9239	0.9339	0.9280	2.249
Facebook	Ada Boost	0.9887	0.9987	0.9874	0.9908	0.9889	**0.124**
	Random Forest	**0.9952**	0.9986	**0.9936**	**0.9970**	**0.9952**	0.521
	XGBoost	0.9887	0.9990	0.9843	0.9938	0.9889	0.366
	Extra Trees	0.9903	**0.9994**	0.9905	0.9908	0.9905	0.468
	LightGBM	0.9920	0.9991	0.9874	**0.9970**	0.9920	0.134
	Gradient Boosting	0.9936	0.9992	0.9905	0.9969	0.9936	0.164
	CatBoost	0.9936	**0.9994**	**0.9936**	0.9939	0.9937	2.807

Table 7. *Selected features by genetic algorithm in Instagram*

	ML classifier						
Features	AdaBoost	RF	XGBoost	Extra Trees	LightGBM	GB	CatBoost
Profilepic	✗	✓	✗	✓	✓	✗	✓
Nums/length username	✓	✓	✓	✓	✗	✓	✗
Fullname words	✓	✓	✓	✓	✓	✓	✗
Nums/length fullname	✗	✓	✗	✗	✓	✗	✓
Name==username	✗	✗	✗	✗	✗	✗	✓
Description length	✓	✓	✓	✓	✓	✓	✗
External URL	✓	✗	✗	✓	✗	✓	✗
Private	✗	✗	✓	✓	✓	✗	✓
Posts	✓	✓	✓	✓	✗	✓	✓
Followers	✓	✓	✓	✓	✓	✓	✓
Follows	✓	✗	✓	✓	✓	✓	✓

features with Gradient Boosting and AdaBoost, that affect the prediction of fake profiles, are: *No Friend, video, checkin, sport, like, group, post shared/post posted rate and about me.*

Table 9 shows the classification results for each classifier examined. AUC is an independent metric for evaluating the performance of a prediction model that relates the True Positive Rate (TPR) to the False Positive Rate (FPR). The best prediction model will have an AUC of 1. According to these results, the performance of the method without feature selection typically exhibited superior performance compared to the one using feature selection methods such as Genetic Algorithm (GA) and Boruta. Nevertheless, it is important to note that the results without feature selection displayed subtle distinctions of less than 07.04% and greater than 1.78% in terms of AUC for all machine-learning classifiers keeping only a subset of 63% among all Instagram features(see Tables 3 and 9). In this context, the findings suggest that a feature selection may eliminate a hundred or even thousands of features. This study also shows an exciting result; Gradient Boosting Classifier got a similar score with Non-selection and GA using 30% of Facebook

Table 8. *Selected features by genetic algorithm in Facebook*

Features	AdaBoost	RF	XGBoost	Extra Trees	LightGBM	GB	CatBoost
				ML classifier			
No friend	✓	✓	✓	✓	✗	✗	✗
Phototag*	✗	✗	✗	✓	✓	✗	✓
Photopost*	✗	✓	✓	✓	✓	✓	✗
Video	✓	✗	✗	✓	✗	✓	✓
Checkin	✓	✓	✓	✓	✓	✓	✓
Sport	✓	✗	✓	✓	✓	✓	✓
Player	✗	✗	✗	✓	✗	✗	✓
Music	✗	✗	✓	✓	✗	✗	✓
Film	✗	✗	✗	✓	✓	✗	✓
Series	✗	✗	✗	✓	✗	✗	✓
Book	✗	✓	✗	✗	✓	✗	✓
Game	✗	✗	✗	✓	✗	✗	✗
Restaurant	✗	✗	✗	✓	✗	✗	✓
Like	✓	✗	✗	✗	✓	✓	✓
Group	✓	✗	✓	✗	✓	✓	✓
Post shared/post posted rate	✓	✗	✓	✓	✓	✓	✓
Education_no	✗	✗	✗	✓	✗	✗	✓
Education_secondary school	✗	✗	✗	✗	✗	✗	✓
Education_university	✗	✗	✗	✗	✓	✗	✓
About me_yes	✓	✗	✓	✓	✓	✓	✗
Family_yes	✗	✓	✗	✗	✗	✗	✗
Gender_male	✗	✓	✗	✓	✓	✗	✓
Relationship_complicate	✗	✗	✗	✓	✓	✗	✓
Relationship_married	✗	✗	✗	✗	✓	✗	✗
Relationship_vide	✗	✗	✗	✓	✗	✗	✗
Note_Yes	✗	✗	✗	✗	✓	✗	✓

features. Moreover, all classifiers (excluding GB Classifier) degraded AUC between 5% and 0.02% when the feature selection was applied to the Facebook dataset.

Machine learning using all characteristics is not always advantageous since it takes considerable time to learn information about many different features. Tables 3, 6, and 9 show the time required to construct a model for each method. Comparing Boruta with Non-selection, the values imply that the genetic algorithm requires less time to construct a model than Boruta.

5. Conclusion and Future Scope

It is crucial to detect fake accounts in OSNs quickly and accurately. Various studies have employed machine learning to detect malicious fake accounts and feature selection to accelerate the process of solving this issue. Experiments were conducted in this study to select the appropriate features to apply machine learning based on previous research, and the findings of this research demonstrate the utility of employing genetic algorithm-based feature selection in comparison to the widely used Boruta technique. Furthermore, this approach offers a notable advantage over non-selection methods, primarily by significantly reducing the time necessary for model construction. However, the feature selection performance using the Gradient Boosting Classifier in conjunction with the GA algorithm yielded a promising result by reducing the number of features by 70%. We can conclude that using data without feature selection is the

Table 9. *Overall results of our experiment using genetic algorithm*

	Classifier	Accuracy	AUC	Recall	Precision	F1-score	Time (s)
Instagram	Ada Boost	0.942	0.942	0.936	0.936	0.936	0.103
	Random Forest	0.9	0.9	0.927	0.876	0.901	0.482
	XGBoost	**0.964**	**0.965**	**0.984**	**0.942**	**0.962**	0.218
	Extra Trees	0.907	0.906	0.880	0.921	0.900	0.114
	LightGBM	0.921	0.920	0.931	0.918	0.925	**0.029**
	Gradient Boosting	0.935	0.934	0.972	0.910	0.940	0.110
	CatBoost	0.935	0.937	0.97	0.902	0.935	0.488
Facebook	Ada Boost	0.991	0.998	0.994	0.994	**0.998**	0.101
	Random Forest	0.983	0.982	0.976	0.987	0.982	0.250
	XGBoost	0.988	0.989	0.979	**0.998**	0.989	0.051
	Extra Trees	0.949	0.950	0.935	0.966	0.950	0.233
	LightGBM	**0.997**	0.997	0.992	0.995	**0.998**	**0.037**
	Gradient Boosting	0.996	**0.999**	**0.998**	0.997	**0.998**	0.137
	CatBoost	0.994	0.995	0.990	0.985	0.995	0.619

optimal method based on the evaluation metrics. However, considering the required time, selecting features using genetic algorithms is preferable. In the future, we aim to contribute to the development of strategies and tools that resonate with the unique dynamics of African OSNs. We also emphasize the importance of adapting machine-learning techniques to take account of the linguistic diversity, users' behavior, and regional trends prevalent in African digital communities.

References

Adewole KS, **Balogun AO**, **Raheem MO**, **Jimoh MK**, **Jimoh RG**, **Mabayoje MA**, **Usman-Hamza FE**, **Akintola AG and Asaju-Gbolagade AW** (2021) Hybrid feature selection framework for sentiment analysis on large corpora. *Jordanian Journal of Computers and Information Technology 7*(2), 1.

Akyon FC and Esat Kalfaoglu M (2019) Instagram fake and automated account detection. In *Proceedings - 2019 Innovations in Intelligent Systems and Applications Conference, ASYU 2019*. Izmir, Turkey: IEEE, pp. 1–7.

Albayati MB and Altamimi AM (2019) Identifying fake facebook profiles using data mining techniques. *Journal of ICT Research and Applications 13*(2), 107–117.

Allam M and Nandhini M (2018) Optimal feature selection using binary teaching learning based optimization algorithm. *Journal of King Saud University-Computer and Information Sciences 34*(2), 329–341.

Anand N, **Sehgal R**, **Anand S and Kaushik A** (2021) Feature selection on educational data using boruta algorithm. *International Journal of Computational Intelligence Studies 10*(1), 27–35.

Bakhshandeh B (2019) Instagram fake spammer genuine accounts.

Bhattasali T and Saeed K (2021) Typing pattern analysis for fake profile detection in social media. In *Computer Information Systems and Industrial Management: 20th International Conference, CISIM 2021, Elk, Poland, September 24–26, 2021, Proceedings 20, Vol. 12883 LNCS*. Cham: Springer International Publishing, pp. 17–27.

Cai J, **Luo J**, **Wang S and Yang S** (2018) Feature selection in machine learning: A new perspective. *Neurocomputing 300*, 70–79.

Carmi E (2020) Rhythmedia: A study of Facebook immune system. *Theory, Culture and Society 37*(5), 119–138.

Deepa S (2008) *Introduction to Genetic Algorithms*. Berlin: Springer.

Deng X, **Li Y**, **Weng J and Zhang J** (2019) Feature selection for text classification: A review. *Multimedia Tools and Applications 78* (3), 3797–3816.

Fraser A, **Burnell D**, *et al.* (1970) *Computer Models in Genetics*, New York: Mcgraw-Hill Book Co.

Galal S (2023) Social media penetration in Africa in 2023, by region. Available at https://www.statista.com/statistics/1190628/social-media-penetration-in-africa-by-region/.

Gazeloglu C (2020) Prediction of heart disease by classifying with feature selection and machine learning methods. *Progress in Nutrition 22*(2), 660–670.

Ghatasheh N, **Altaharwa I and Aldebei K** (2022) Modified genetic algorithm for feature selection and hyper parameter optimization: Case of XGBoost in spam prediction. *IEEE Access 10*, 84365–84383.

Gupta A and Kaushal R (2017) Towards detecting fake user accounts in facebook. In *ISEA Asia Security and Privacy Conference 2017, ISEASP 2017*, Vol. *1*. Surat, India: IEEE, pp. 1–6.

Jennings PC, **Lysgaard S**, **Hummelshøj JS**, **Vegge T and Bligaard T** (2019) Genetic algorithms for computational materials discovery accelerated by machine learning. *npj Computational Materials 5*(1), 46.

Joshi S, **Nagariya HG**, **Dhanotiya N and Jain S** (2020) Identifying fake profile in online social network: An overview and survey. In *International Conference on Machine Learning, Image Processing, Network Security and Data Sciences*. Singapore: Springer, pp. 17–28.

Katoch S, **Chauhan SS and Kumar V** (2021) A review on genetic algorithm: Past, present, and future. *Multimedia Tools and Applications 80*(5), 8091–8126.

Kaubiyal J and Jain AK (2019) A feature based approach to detect fake profiles in twitter. In *ACM International Conference Proceeding Series*. New York, NY: Association for Computing Machinery, pp. 135–139.

Kursa M and Rudnicki W (2020) Boruta: Wrapper algorithm for all relevant feature selection. *Visité le 6*(10), 2020.

Liu X-Y, **Liang Y**, **Wang S**, **Yang Z-Y and Ye H-S** (2018) A hybrid genetic algorithm with wrapper-embedded approaches for feature selection. *IEEE Access 6*, 22863–22874.

Ma W, **Zhou X**, **Zhu H**, **Li L and Jiao L** (2021) A two-stage hybrid ant colony optimization for high-dimensional feature selection. *Pattern Recognition 116*, 107933.

Mohammadrezaei M, **Shiri ME and Rahmani AM** (2018) Identifying fake accounts on social networks based on graph analysis and classification algorithms. *Security and Communication Networks 2018*, 1–8.

Moslehi F and Haeri A (2020) A novel hybrid wrapper–filter approach based on genetic algorithm, particle swarm optimization for feature subset selection. *Journal of Ambient Intelligence and Humanized Computing 11*(3), 1105–1127.

Munga JB and Mohandas P (2022) Feature selection for identification of fake profiles on facebook. In Usman J, Liew YM, Ahmad MY and Ibrahim F (eds), *6th Kuala Lumpur International Conference on Biomedical Engineering 2021*. Cham: Springer International Publishing, pp. 489–497.

Nettleton D (2014) Selection of variables and factor derivation. In *Commercial Data Mining*, Boston: Morgan Kaufmann, pp. 79–104.

Shami TM, **El-Saleh AA**, **Alswaitti M**, **Al-Tashi Q**, **Summakieh MA and Mirjalili S** (2022) Particle swarm optimization: A comprehensive survey. *IEEE Access 10*, 10031–10061.

Rácz A, **Bajusz D and Héberger K** (2019) Multi-level comparison of machine learning classifiers and their performance metrics. *Molecules 24*(15), 2811.

Rostami M, **Berahmand K and Forouzandeh S** (2021) A novel community detection based genetic algorithm for feature selection. *Journal of Big Data 8*(1), 1–27.

Sahoo SR and Gupta BB (2020) Classification of spammer and nonspammer content in online social network using genetic algorithm-based feature selection. *Enterprise Information Systems 14*(5), 710–736.

Sheikhi S (2020) An efficient method for detection of fake accounts on the instagram platform. *Revue d'Intelligence Artificielle 34*(4), 429–436.

Singhal Y, **Jain A**, **Batra S**, **Varshney Y and Rathi M** (2018) Review of bagging and boosting classification performance on unbalanced binary classification. In *2018 IEEE 8th International Advance Computing Conference (IACC)*. Greater Noida, India: IEEE, pp. 338–343.

Venkatesh B and Anuradha J (2019) A review of feature selection and its methods. *Cybernetics and Information Technologies 19*(1), 3–26.

Wang X, **Lai CM**, **Lin YC**, **Hsieh CJ**, **Wu SF and Cam H** (2019) Multiple accounts detection on facebook using semi-supervised learning on graphs. In *Proceedings - IEEE Military Communications Conference MILCOM*. Los Angeles, CA: IEEE, pp. 1–9.

Wani MA, **Agarwal N**, **Jabin S and Hussain SZ** (2019) Analyzing real and fake users in facebook network based on emotions. In *2019 11th International Conference on Communication Systems and Networks, COMSNETS 2019*, Vol. *2061*. Bengaluru, India: IEEE, pp. 110–117.