Bio-inspired algorithms for effective social media profile authenticity verification

Nadir Mahammed[1], Badia Klouche[1], Imène Saidi[1], Miloud Khaldi[1] and Mahmoud Fahsi[2]

[1]LabRI-SBA Laboratory, Ecole Superieure en Informatique Sidi Bel Abbes, P.O 73, El Wiam Sidi Bel Abbés 22016, Algeria
[2]EEDIS Laboratory, Djillali Liabes University, P.O 89 Sidi Bel Abbès 22000, Algeria

Abstract

In the ever-evolving digital era, the profound impact of online social networks is omnipresent. Platforms like Instagram, Facebook, and Twitter grapple persistently with the challenge of distinguishing genuine user profiles from a rising tide of counterfeit or dormant accounts. This predicament underscores the critical need to adeptly differentiate between authentic and misleading user profiles, particularly in light of the increasing prevalence of online deception. This research centers on introducing an innovative approach to profile validation, highlighting the pivotal task of identifying and mitigating the presence of fake profiles across social media platforms. The methodology employed is groundbreaking, strategically integrating cutting-edge bio-inspired algorithms, with a specific emphasis on the application of metaheuristics. Unlike conventional machine learning techniques, this approach navigates the intricate landscape of online social networks with unparalleled agility and adaptability. Despite the inherent challenges posed by the nature and scarcity of datasets available on the web, the empirical results are remarkably compelling. The approach consistently demonstrates a high level of accuracy in classification tests, showcasing its efficacy in addressing the pervasive issue of fake profiles in the digital realm.

Keywords

Social media, fake profile detection, bio-inspired algorithm, machine learning, simulation

1. Introduction

In the ever-evolving landscape of online social networks, as exemplified by the behemoths Facebook and Twitter, a remarkable surge in user engagement has occurred over recent years. This rapid growth, however, has been accompanied by a troubling escalation in the presence of fake accounts and online impersonation. This issue is not only on the rise but has also gained significant scholarly attention, as evident in [1] report on detecting fake profiles. The essence of these fake profiles lies in their representation of fictitious personas or entities that expertly mimic real users, raising pertinent concerns within the online social network ecosystem.

One of the fundamental challenges in this domain is the absence of robust authentication mechanisms on many social networking platforms. These mechanisms are instrumental in effectively distinguishing between genuine user accounts and fraudulent counterparts. As underscored by [2]. in their 2022 survey, the deficiencies in these mechanisms exacerbate the proliferation of fake accounts, thus prompting a dire need for an innovative and effective solution. Such a solution is essential to identify and mitigate the presence of counterfeit accounts, ultimately ensuring the creation of a secure and trustworthy environment for the multitude of users frequenting social networking sites.

In addressing this pressing concern, the authors of this study have embarked on a transformative journey, departing from the well-trodden path of Machine Learning (ML) methods to explore the promising realm of metaheuristics. Within this domain, they have harnessed the capabilities of the Fire Hawk Optimizer (FHO), a contemporary bio-inspired algorithm, to address the multifaceted challenge of fake profile detection. This unconventional approach represents a noteworthy departure from conventional methodologies and stands as a beacon of innovation, poised to revolutionize the field of online social network analysis.

The ensuing sections of this comprehensive study delve into the foundational principles and practical implications of this pioneering approach. By elucidating its diverse facets, the study aims to underscore the transformative potential of FHO in the context of enhancing the security and authenticity of online social networks on a global scale. Thus, it transcends mere theoretical exploration and emerges as a promising catalyst for substantive change in the landscape of social network analysis and the broader digital sphere.

6th International Hybrid Conference On Informatics And Applied Mathematics, December 6-7, 2023 Guelma, Algeria
*Nadir Mahammed
n.mahammed@esi-sba.dz (N. Mahammed); b.klouche@esi-sba.dz (B. Klouche); i.saidi@esi-sba.dz (I. Saidi); m.khladi@esi-sba.dz (M. Khaldi); mahmoud.fahsi@univ-sba.dz (M. Fahsi)
0000-0001-7865-5937 (N. Mahammed); 0000-0001-7417-612X (I. Saidi); 00000022896136X (M. Fahsi)

2. Related Work

This section undertakes a thorough examination of recent developments in detecting fake profiles within the domain of online social media. This investigation goes beyond the limitations of traditional machine learning techniques, illuminating a multifaceted landscape of innovative approaches and methodologies.

In addressing the escalating problem of fake profiles on prominent online social networks such as Facebook and Twitter, Mahammed et al. [3] propose a pioneering solution. The paper introduces an innovative approach to detect fake profiles by amalgamating a machine learning algorithm with a bio-inspired algorithm. This hybrid methodology comprises two stages, with the initial stage leveraging the Satin Bowerbird Optimization algorithm (SBO) to identify the optimal initial centroid for the subsequent k-means clustering algorithm in the second stage. The effectiveness of this approach surpasses that of established machine learning algorithms in the realm of fake profile detection, underscoring the paramount significance of ensuring authenticity and security in online social network interactions.

The authors of [4] tackle the issue of deceptive online reviews and ratings that can mislead consumers in their purchasing decisions. They introduce two innovative deep-learning hybrid techniques: CNN-LSTM for identifying fake online reviews and LSTM-RNN for detecting fraudulent ratings. These models surpass existing methods and achieve high prediction accuracy, particularly when applied to Amazon datasets.

In [5], the primary focus is on addressing privacy concerns and catering to brands and marketers. The proposed enhancements include expanding the prototype, enabling bulk image processing, integrating with social media APIs, improving the model's recall for logo detection, increasing generalizability, exploring logo localization, combining textual analytics, automating model selection and hyper-parameter tuning, and comparing performance with existing logo detection systems.

The research conducted by [6] introduces the Multi-Relational Graph-Based Twitter Account Detection Benchmark (MGTAB) to advance social media user stance and bot detection methods. MGTAB overcomes issues of low annotation quality and incomplete user relationships present in existing benchmarks by providing a comprehensive dataset comprising 1.55 million users and 130 million tweets.

[7] develops a method for identifying and verifying duplicate profiles in online social networks. The approach involves using attribute and network-based similarity measures, implementing the model with MapReduce to reduce computational complexity, and creating a testing dataset. The study employs parallel k-means clustering and parallel SVM classification techniques to effectively identify suspicious profiles within clusters containing genuine ones.

The work by [8] introduces a novel technique named GWODS for detecting attacker shilling profiles in recommender systems. GWODS combines the K-means clustering algorithm with the Grey Wolf Optimizer (GWO) to identify suspicious profiles. It demonstrates promising results in experiments conducted on MovieLens datasets and can be employed as a preprocessing step to prevent biased recommendations in recommender systems.

In the article by [9], the core concept revolves around the limitations faced in social media bot detection, particularly on Twitter, regarding data collection methods. While machine learning-based algorithms exhibit near-perfect performance on existing datasets, the study reveals that accuracy is often influenced by dataset-specific factors rather than inherent differences between humans and bots. Additionally, the use of decision trees is preferred due to their interpretability over random forest classifiers.

Table 1 provides a comprehensive overview of the state of research in fake profile detection, emphasizing the need for further investigations that integrate various techniques, improve generalization, and address the dynamic nature of online threats in OSNs. Fake profile detection is a critical aspect of maintaining the integrity and security of online platforms, and these studies play a crucial role in advancing the field.

- Diverse Research Efforts: The table underscores a broad spectrum of research initiatives aimed at fake profile detection, indicating a heightened awareness of the severity of fake profiles in Online Social Networks (OSNs) and the urgency to address this issue. This diversity suggests multiple avenues being explored to tackle the problem.
- OSN-Specific Approaches: Several studies focus on specific OSNs like Facebook, Instagram, and Twitter, acknowledging the unique characteristics and challenges of each platform. This prompts the question of whether a universal model can effectively detect fake profiles across various OSNs or if tailored solutions are necessary.
- Machine Learning and Metaheuristics: Utilized techniques range from traditional machine learning algorithms (Decision Trees, Random Forest, Support Vector Machine, and K-means) to bio-inspired metaheuristics (Satin Bowerbird Optimization and Grey Wolf Optimizer). This mix indicates exploration of both data-driven and heuristic-driven approaches, warranting research into their relative efficacy and optimal use.
- Incorporation of Deep Learning: Some studies incorporate deep learning methods, such as Con-

Table 1
Related work summary

Reference	OSN	ML	Metaheuristic	Other	Dataset	Results (acc)
[3]	Facebook	SVM,NB,RF,KNN	SBO	-	1244	0.98
[4]	-	-	-	CNN,LSTM,RNN	20000	0.93
[5]	Instagram	-	-	CDS	10000	0.90
[6]	Twitter	RF,DT,SVM	-	Adaboost	130 millions	0.97
[7]	Facebook	SVM, K-means	-	MapReduce	1000	0.98
[8]	-	k-means	GWO	-	6000	0.99
[9]	Twitter	DT,RF	-	-	-	0.91

volutional Neural Networks, Long Short-Term Memory, and Recurrent Neural Networks, highlighting the need for advanced methods to combat sophisticated fake profiles employing deep learning in their creation.

- Dataset Size and Quality: Dataset size plays a pivotal role, with some studies employing datasets containing millions of instances. While larger datasets offer more robust training, they also demand greater computational resources. Additionally, dataset quality is crucial, necessitating research into effective collection and curation techniques.
- Accuracy Achievements: Notably, some studies achieve very high accuracy levels (e.g., 0.98 and 0.99). While promising, it's vital to scrutinize the generalization capabilities of these models, as high accuracy on one dataset doesn't guarantee success on new, unseen data.
- Challenges and Future Directions: Challenges include the evolving techniques in fake profile creation and the need for real-time or near-real-time detection. Future research should address these challenges and explore methods for dynamic model adaptation.
- Integration and Model Ensemble: Combining strengths from different models or creating ensemble models can potentially enhance detection accuracy. Research in this direction could lead to more robust solutions.
- Explainability and Interpretability: As fake profile detection systems are deployed, there's a growing need for interpretability and explainability in model decisions, especially in legal and ethical contexts.
- Scalability: Ensuring scalability of fake profile detection methods to handle the increasing volume of data on OSNs is a significant concern. Research should focus on algorithm efficiency in large-scale scenarios.

From this bibliographic study, it is deduced that employing metaheuristics for detecting fake profiles on social networks proves to be a crucial approach. These optimization methods offer notable advantages in terms of efficiency, computation time, and resilience to data variations—key elements in the field of fake profile detection on social networks. Metaheuristics excel in effectively exploring solution spaces, adapting well to complex landscapes. This enhanced exploration capability enables convergence toward high-quality solutions, even in poorly defined search spaces. Moreover, metaheuristics are recognized for their computational efficiency, often converging to acceptable solutions within reasonable timeframes, making them particularly well-suited for complex problems. Furthermore, they exhibit robustness in the face of data variations, requiring less dependence on the specific nature of the data and demonstrating adaptability to incomplete or noisy datasets.

From this bibliographic study, it is deduced that employing metaheuristics for detecting fake profiles on social networks proves to be a promising approach for addressing challenges in artificial intelligence and machine learning in this specific domain, offering high-quality solutions, optimized computation time, and independence from data variations.

3. Material and Methodology

3.1. Dataset

Employing distinct batches for labeling, the dataset construction involved the first batch, which comprised Twitter data sourced from previously banned pro-ISIS accounts, serving as positive labels. Specifically, the dataset "How ISIS Uses Twitter" was utilized [1], encompassing 17,350 tweets from over 110 pro-ISIS accounts. This dataset includes attributes (see table 2 such as Name, Username, Description, Location, Number of followers at the time of tweet download, Number of statuses by the user when the tweet was downloaded, Date and timestamp of the tweet, and the tweet itself. To address Arabic content, the Google Translate API was utilized for translation.

[1] https://www.kaggle.com/fifthtribe/how-isisuses-twitter

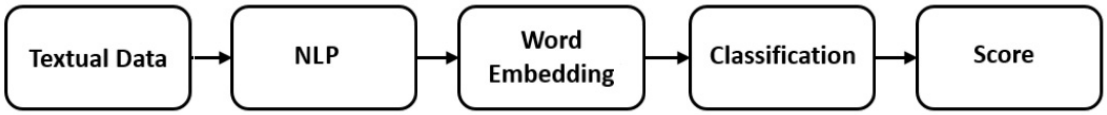

Figure 1: Text classification method.

Table 2
Dataset description

Attribute	Defintion
Name	Name of the user
Username	User's Twitter username
Description	User's profile description
Location	User's specified location
Followers	Nbr of followers at tweet download
Statuses	Nbr of user's statuses at tweet download
Timestamp	Date and timestamp of the tweet
Tweet	The content of the tweet

For the second batch, the Global Terrorism Database (GTD) was employed as a negative labeled dataset [10] [11] . The GTD contains information on over 180,000 terrorist attacks worldwide since 1970. Filtering events from 2002 onwards, data was extracted from the "summary" column, which provides summaries of each attack.

3.2. Text classification

The process of discerning information from textual input involves three principal stages, as depicted in Figure 1.

- Natural Language Processing (NLP): This initial phase focuses on preprocessing textual data, ensuring a well-structured format for ease of understanding and processing. The analysis of textual data unfolds in four essential steps: tagging, annotating, co-reference resolution, and sentiment analysis [12].
- Word Embedding : Embracing the N-gram language model [13], the probability is estimated of the last word based on preceding words. This choice is informed by its superior performance compared to the TF-IDF model [14].
- Classification: Post word embedding, the textual content takes on a numerical form, making it machine-readable. This numerical representation is then input into a classifier, allowing the model to effectively perform the classification task.

3.3. Preprocessing

Data preprocessing is the process of converting raw data into a format that can be readily understood by machine learning algorithms. As detailed in [15], the data preparation procedures for the different datasets employed in this research are succinctly outlined below:

1. Data Scrutiny: Eliminate duplications and rectify errors.
    a) Eliminate duplications, superfluous data points, inaccuracies, and redundant columns (such as 'id' and 'id-name').
    b) Omit irrelevant data points, inaccuracies, and redundant columns (such as 'id' and 'id-name').
2. Address disparities, anomalies, and missing data.
3. Standardize and adapt the data through scaling.
4. Prune interrelated variables and streamline the dataset.

3.4. Machine Learning Algorithms

3.4.1. Induction of Decision Tree

When considering decision tree induction, it is noteworthy that ID3 operates as a supervised learning algorithm. This method constructs a tree based on information derived from training instances, utilizing it for classifying test data [16].

3.4.2. K-means Algorithm

A cornerstone in unsupervised learning for pattern recognition and machine learning, the K-means algorithm is renowned for its simplicity and widespread use among iterative and hill-climbing clustering algorithms [17].

3.4.3. Hierarchical Clustering Analysis

Hierarchical clustering (HC) groups similar objects into clusters. Starting with each object as a separate cluster, it iteratively merges the closest clusters until forming a single, hierarchical structure. This method is valuable for revealing data patterns and relationships [18].

3.4.4. Nearest Neighbor Classification

Often referred to as K-nearest neighbors (KNN), this method is grounded in the concept that the nearest patterns to a target pattern, for which a label is sought, offer valuable label information [**?**].

3.4.5. Naive Bayes Classifier

Commonly known as NB, the Naive Bayes classifier is a supervised learning algorithm rooted in Bayes' theorem. It operates on the simplifying assumption that attribute values are conditionally independent when considering the target value [19].

3.4.6. Random Forest Machine

Random forests (RF) represent an amalgamation of tree predictors. Each tree relies on the values of a random vector, independently sampled with a uniform distribution shared across all trees within the forest [20].

3.4.7. Support Vector Machine

The Support Vector Machine (SVM) is recognized as a potent tool for classifier construction. SVM is purposefully designed to establish a robust decision boundary between two classes, facilitating the accurate prediction of labels from one or more feature vectors [21].

3.5. Proposed Algorithm

3.5.1. Inspiration

Australia's Indigenous people have a rich history of employing fire as a tool for ecosystem management. Controlled burns, whether ignited intentionally or by lightning, play a crucial role in maintaining the balance of the environment. However, a fascinating revelation involves certain bird species, known as Fire Hawks, which include whistling kites, black kites, and brown falcons. These birds have been observed intentionally carrying burning sticks and using them to start fires as part of their predatory tactics. This behavior is strategic, as the induced fires serve to startle and capture prey such as rodents, snakes, and other animals, enhancing the efficiency of their hunting endeavors.

3.5.2. Motivation to choose

This nature-inspired strategy, finely tuned over eons of evolution, equips the Fire Hawk Optimizer (FHO) for intricate optimization tasks. FHO excels in rapid convergence, surpassing alternative methods. Its robust nature allows effective handling of noisy and uncertain data, contributing to enhanced solution exploration diversity.

The remarkable convergence speed of FHO is valuable in time-sensitive or resource-constrained scenarios. It swiftly reaches optimal solutions through iterations until predefined criteria are met. FHO's computational efficiency is evident as it converges to the global optimum with fewer evaluations [22].

3.5.3. Operation

The FHO algorithm, inspired by the foraging behavior of fire hawks, operates through the following steps:

1. Initial Positioning: At the start, solution candidates (X) are defined, representing the positions of fire hawks and prey in the search space. Random initialization places these vectors within the search space, taking into account various parameters.
2. Fire Hawks and Prey: The algorithm categorizes solution candidates into Fire Hawks and prey based on their objective function values. Selected Fire Hawks aim to spread fires around the prey, with the global best solution serving as the primary fire source.
3. Determining Territories: The algorithm calculates the total distance between Fire Hawks and prey to identify the nearest prey to each bird. This step determines the effective territory of the Fire Hawks for hunting. The bird with the best objective function value selects the nearest prey to its territory, while others choose their next nearest prey.
4. Spreading Fires: Fire Hawks collect burning sticks from the main fire and drop them in their territories, causing the prey to flee. Some Fire Hawks may use burning sticks from other territories, contributing to position updates in the search loop.
5. Prey Movements: The prey's movements within Fire Hawks' territories are considered. The algorithm simulates various prey actions, such as hiding, running, or approaching Fire Hawks, impacting position updates.
6. Safe Places: Prey may move toward safe places outside Fire Hawk territories. These movements are also included in the position update process.
7. Territory Definition: Fire Hawk territories are represented as circular areas, with the precise territory determined by prey numbers and distances from each Fire Hawk.
8. Boundary Violation and Termination: The algorithm considers boundary control for violating decision variables and employs a termination criterion, such as a predefined number of objective function evaluations or iterations, to conclude the process.

The figure 2 provides pseudocode which offers a concise overview of the FHO algorithm's operation.

3.5.4. Transition from natural to artificial

This section is devoted to examining the shift from the Fire Hawk's innate behaviors in the wild to its adapted

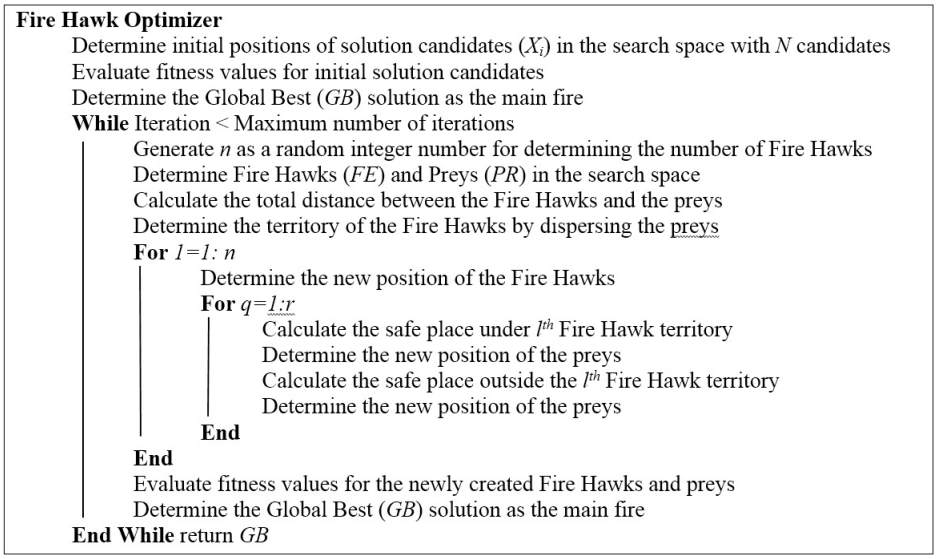

Figure 2: FHO Pseudocode.

Table 3
Transition from natural to artificial of FHO

Natural	Artificial
Natural Artificial Fire hawks hunting for prey in the wild	Each user is classified into the most Suitable class (Real or Fake)
Fire hawks finding food by following the smoke signals from wildfires	Two classes (Real or Fake)
Environment	Online Social Networks (Facebook, Twitter, Instagram)
Fire hawk	Online Social Networks User
Group of fire hawks	Online Social Networks Users
Best individual in the group of fire hassle that found the prey	The best solution in the population of solutions that meets the objective function by the FHO (Real or Fake)
The distance between the flee hawk and its prey	$D_k^1 = \sqrt{(x_2 - x_1) + (y_2 - y_1)}$ The distance between the solution and the optimal

behaviors in an artificial environment, as detailed in the table 4.

Table 4 delves into a captivating comparison between the natural and artificial, spotlighting the FHO algorithm's mission of distinguishing genuine from fraudulent profiles in online social networks. It intriguingly parallels the hunting behavior of fire hawks with user suitability assessment.

By mentioning distance calculations, it hints at the algorithm's quest for the optimal solution, equating to precise user classifications in social networks. This table is a gateway to understanding how nature's wisdom inspires advanced algorithms that address real-world challenges.

It embodies the fusion of the natural and artificial realms, demonstrating how algorithmic innovation stems from nature's timeless principles, resolving complex issues in online social networks. Ultimately, it invites exploration of the limitless possibilities born from the fusion of nature and algorithms.

3.5.5. Fitness function

The FHO rigorously employs a fitness function, as depicted in Figure 3, to meticulously gauge the performance of solution candidates. This fitness function pivots around the precision of a gradient boosting classifier meticulously applied to a thoughtfully selected subset of features sourced from a dataset.

To elaborate on the computation of the fitness value, the function takes a solution candidate into its fold, representing a distinct subset of features. This subset undergoes scrupulous evaluation via a gradient boosting

Fitness Function

```
def fitness function(solution):
classifier = GradientBoostingClassifier(n estimators=100, random state=42)
selector = SelectFromModel(classifier)
selector.fit(X, y)
X selected = selector.transform(X)
classifier.fit(X selected, y) ▷Return the accuracy score on the training set
return accuracy score(y, classifier.predict(X selected))
```

Figure 3: Fitness Function.

classifier, armed with precisely 100 estimators and a deterministic random state fixed at 42. Notably, this classifier undertakes the dual responsibility of feature selection and classification.

The inner workings of the fitness function encompass the formulation of a feature selector. This selector, entailing sophisticated intricacies, leverages the classifier itself to discern and pinpoint the paramount features based on the classifier's predictive capabilities. This discernment is crucial in optimizing the classification process.

Of particular significance is the selector's subsequent fitting to both the input dataset and the target variable. This preparatory phase is pivotal for the forthcoming accuracy evaluation.

What distinguishes this fitness function is its intrinsic capacity to bring about a transformation of the input dataset. This transformation is rendered by carefully cherry-picking the most pivotal features from the original dataset. The result is a transformed dataset, which bears the promise of enhanced accuracy. This transformed dataset now becomes the testing ground for the classifier. It serves as the substrate for the classifier's extensive training process, conducted in close tandem with the target variable.

As the final step in this intricate dance of precision, the fitness function introduces the crucial concept of the accuracy score. It orchestrates a meticulous comparison between the true labels and the predicted labels that emerge from the classifier's outputs on the transformed dataset. The resultant accuracy score stands as a testament to the chosen subset of features' ability to effectively forecast the target variable.

Figure 4 demonstrates the pivotal role of the fitness function in the FHO. In the third stage of the code, the fitness values for each solution candidate in the population are meticulously computed by invoking the fitness function. This function is systematically applied to every row (axis=1) within the population array, yielding an array replete with fitness values, which are more specifically accuracy scores. These accuracy scores bear significance as they provide a quantitative assessment of each solution candidate's performance accuracy.

Table 4
FHO metrics

Parameter	Value
Population size	50
Iteration	100
Iteration by dataset	100

In essence, the fitness function operates as the core evaluator, discerning and ranking solution candidates based on their individual performance. In the broader context, these fitness scores wield substantial influence in steering the FHO's pursuit of the optimal solution, with the overarching goal of optimizing performance accuracy.

3.5.6. FHO metrics

The FHO algorithm undergoes a comparative analysis against a spectrum of established Machine Learning algorithms, encompassing ID3, SVM, NB, RF, HC, KNN with diverse K values, and K-means. This exhaustive evaluation consists of 100 iterations for each dataset, ensuring robustness and careful examination. Notably, the FHO configuration parameters are as follows: the initial population size is set at 50, and the maximum number of iterations is capped at 100 as summarized in Table 4

4. RESULTS AND DISCUSSION

Throughout the experimental phase, a 2014 MSI GT70 gaming laptop was employed, featuring an Intel Core i7-4800MQ CPU, a Nvidia GeForce GTX 770M GPU, and 32 GB of RAM.

Fitness Function within FHO

```
# Determine the search space and initialize solution.
1: search space = (0, X.shape[1]) candidates
2: population = np.random.rand(self.population size, X.shape[1])
# Evaluate fitness values for initial solution candidates.
3: fitness values = np.apply along axis(fitness function, 1, population)
```

Figure 4: Fitness Function with FHO.

Table 5
Results classes

Class	Meaning
True Positives (TP)	Instances where the model accurately identifies positive cases within the dataset.
False Positives (FP)	Cases in which the model incorrectly categorizes negative instances as positives.
True Negatives (TN)	Cases in which the model correctly recognizes negative instances.
False Negatives (FN)	Instances in which the model erroneously categorizes positive cases as negative.

4.1. Evaluation Criteria

The detection of fake accounts can be evaluated using various performance metrics, such as Accuracy, F-score, Recall, precision, and entropy. These metrics provide insights into the model's performance and its ability to classify profiles correctly.

In addition, the Confusion Matrix is used as a visual representation of fake account detection, offering a comprehensive view of the model's performance across different classes as shown in Table 5.

- Accuracy: This metric measures the overall accuracy of the model in correctly classifying profiles.

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

- Precision: Calculates the model's accuracy in classifying values correctly by comparing the number of accurately classified profiles to the total classified data points for a given class label.

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

- Recall: This metric assesses the model's ability to correctly predict positive values, indicating how often it correctly identifies true positives.

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

- F1-score: Which is the harmonic mean of precision and recall, balances the trade-off between these two metrics.

$$F1 - score = \frac{2 * TP}{2 * TP + FP + FN} \quad (3)$$

- Entropy: This metric quantifies the randomness or disorder in a system, providing valuable information about the data's structure and organization.

$$Entropy = log2(Precision) * (-Precision) \quad (4)$$

4.2. Results

Table 6 summarizes the obtained results in comparison to the original work conducted with the same dataset [23]. So, the results presented in Table 6 showcase the performance metrics of various classifiers, with a particular emphasis on the Fire Hawk Optimizer (FHO).

FHO stands out prominently, achieving remarkable accuracy, precision, recall, and F1-score values of 99.6%. This outstanding performance suggests that FHO excels in accurately classifying instances, achieving an almost perfect balance between precision and recall. Such high metrics underscore the effectiveness of FHO in the given classification task, highlighting its potential as a robust optimization algorithm.

Comparatively, traditional machine learning classifiers, such as Support Vector Machine (SVM), Naive Bayes

Table 6
Obtained results.

	Accuracy	Precision	Recall	F1-score
FHO	**0.996**	**0.996**	**0.997**	**0.996**
SVM	0.907	0.907	0.902	0.904
NB	0.904	0.904	0.899	0.900
LR	0.899	0.899	0.854	0.875

(NB), and Logistic Regression (LR), demonstrate competitive yet comparatively lower performance. SVM, while achieving a respectable accuracy of 90.7%, falls short of FHO's exceptional accuracy. Similarly, NB and LR, with accuracies of 90.4% and 89.9

Precision, recall, and F1-score values further emphasize FHO's dominance, outperforming the other classifiers across all metrics. The precision of 99.6% indicates an incredibly low false positive rate, essential for tasks where misclassification has significant consequences. The recall of 99.7% highlights FHO's ability to capture the majority of actual positives. The F1-score of 99.6% reflects the harmonious balance between precision and recall.

The outstanding performance of FHO positions it as a formidable tool for classification tasks. Its ability to achieve near-perfect accuracy and balance between precision and recall showcases its potential to outshine traditional machine learning methods in complex optimization scenarios. This reaffirms the significance of bio-inspired algorithms, like FHO, in pushing the boundaries of optimization and classification tasks.

4.3. Discussion

FHO's standout attribute is its remarkable ability to rapidly converge towards predefined tolerance for the global best solution. This swift convergence, coupled with its resource-efficiency, assumes particular significance in the context of social networks where timely profile verification is crucial, and computational resources often come at a premium.

What sets FHO apart is its innate knack for handling the unpredictability and noise inherent in real-world data, showcasing its robustness and adaptability in navigating the often erratic nature of user-generated profile information.

FHO's penchant for diversifying the search process, inspired by natural systems, is another remarkable trait. By concurrently exploriFHO stands out due to its remarkable capacity to swiftly converge toward a predefined tolerance for the globally optimal solution. This rapid convergence, coupled with its resource-efficient nature, holds particular significance in the realm of social networks, where timely profile verification is imperative, and computational resources are often scarce.

What distinguishes FHO is its inherent ability to navigate the unpredictability and noise inherent in real-world data, illustrating its robustness and adaptability in handling the often erratic nature of user-generated profile information.

FHO's inclination to diversify the search process, drawing inspiration from natural systems, is another noteworthy trait. By concurrently exploring multiple potential solutions, it enhances the likelihood of discovering innovative answers, a crucial asset when dealing with the ever-evolving strategies employed by creators of fake profiles.

The results underscore FHO's exceptional computational efficiency, consistently converging to the globally optimal solution within a significantly reduced timeframe. This efficiency proves highly relevant in situations where time sensitivity and the conservation of computational resources are paramount. An additional notable aspect is FHO's ability to converge toward the globally optimal solution in mathematical test functions while requiring fewer objective function evaluations. This underscores its computational efficiency, highlighting its practical applicability across a spectrum of problem-solving scenarios.ng multiple potential solutions, it enhances the likelihood of uncovering innovative answers, a pivotal asset when contending with the ever-evolving strategies employed by creators of fake profiles.

The results speak to FHO's exceptional computational efficiency. It consistently converges to the global best solution within a significantly reduced timeframe, allowing it to swiftly identify optimal or near-optimal solutions. This efficiency proves highly pertinent in situations where time sensitivity and conservation of computational resources are paramount. An additional noteworthy aspect is FHO's ability to converge toward the global best solution in mathematical test functions while requiring fewer objective function evaluations. This underscores its computational efficiency, highlighting its practical applicability across a spectrum of problem-solving scenarios.

5. Conclusion

Within the online social media landscape, the issue of fake profiles has become a prominent concern, particularly on major platforms such as Instagram, Facebook, and Twitter. The widening gap between registered profiles and genuinely active users signals a troubling increase in counterfeit or inactive accounts, posing risks to platform credibility, security, and privacy. Academic literature has predominantly focused on applying machine learning techniques to discern real from fraudulent profiles by analyzing various attributes and user behav-

ior patterns. However, these traditional methods exhibit limitations, prompting the exploration of more robust and efficient solutions.

A transformative shift in the fight against fake profiles has emerged, emphasizing the potential of metaheuristic algorithms, specifically bio-inspired algorithms. This shift acknowledges the constraints of conventional machine learning in handling the complexities of online social network data. Bio-inspired algorithms, exemplified by the Fire Hawk Optimizer (FHO), have shown promise in fake profile detection, deriving computational prowess from their inherent bio-inspired nature, drawing inspiration from the foraging behavior of fire hawks.

The metaheuristic aspect of FHO enhances its significance. As a member of the metaheuristics family, FHO belongs to a class of optimization algorithms praised for their adaptability and efficiency. FHO distinguishes itself by pursuing diverse solution candidates, making it adept at addressing multifaceted challenges, particularly in fake profile detection.

FHO's proficiency is evident in performance results with Instagram, Facebook, and Twitter datasets. It excels in promptly and efficiently converging toward the global best solution, a crucial trait in scenarios where timely profile validation and limited computational resources are critical. Its resilience in handling unpredictable data and its ability to diversify the search process are valuable assets when confronting the evolving tactics of fake profile creators. Furthermore, its computational efficiency, marked by a lower number of objective function evaluations while consistently converging to the global best solution, positions it as a computational prowess exemplar.

Looking ahead, refining and advancing FHO's capabilities for large datasets with heterogeneous data could be a future perspective. Integrating FHO with other advanced techniques and exploring hybrid approaches that leverage its strengths alongside complementary methods for even more robust profile validation are compelling avenues for future studies.

References

[1] R. Bhambulkar, S. Choudhary, A. Pimpalkar, Detecting fake profiles on social networks: A systematic investigation, in: 2023 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), IEEE, 2023, pp. 1–6.

[2] J. Shamseddine, M. Malli, H. Hazimeh, Survey on fake accounts detection algorithms on online social networks, in: The International Conference on Innovations in Computing Research, Springer, 2022, pp. 375–380.

[3] N. Mahammed, S. Bennabi, M. Fahsi, B. Klouche, N. Elouali, C. Bouhadra, Fake profiles identification on social networks with bio inspired algorithm, in: 2022 First International Conference on Big Data, IoT, Web Intelligence and Applications (BIWA), IEEE, 2022, pp. 48–52.

[4] N. Deshai, B. B. Rao, et al., Deep learning hybrid approaches to detect fake reviews and ratings, Journal of Scientific & Industrial Research 82 (2022) 120–127.

[5] V. Tanniru, T. Bhattacharya, Online fake logo detection system (2023).

[6] S. Shi, K. Qiao, J. Chen, S. Yang, J. Yang, B. Song, L. Wang, B. Yan, Mgtab: A multi-relational graph-based twitter account detection benchmark, arXiv preprint arXiv:2301.01123 (2023).

[7] A. Saravanan, V. Venugopal, Detection and verification of cloned profiles in online social networks using mapreduce based clustering and classification, International Journal of Intelligent Systems and Applications in Engineering 11 (2023) 195–207.

[8] S. Bansal, N. Baliyan, Detecting group shilling profiles in recommender systems: A hybrid clustering and grey wolf optimizer technique, in: Design and Applications of Nature Inspired Optimization: Contribution of Women Leaders in the Field, Springer, 2023, pp. 133–161.

[9] C. Hays, Z. Schutzman, M. Raghavan, E. Walk, P. Zimmer, Simplistic collection and labeling practices limit the utility of benchmark datasets for twitter bot detection, in: Proceedings of the ACM Web Conference 2023, 2023, pp. 3660–3669.

[10] G. LaFree, L. Dugan, Introducing the global terrorism database, Terrorism and political violence 19 (2007) 181–204.

[11] J. Lutz, B. Lutz, Global terrorism, Routledge, 2019.

[12] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, P. Kuksa, Natural language processing (almost) from scratch, Journal of machine learning research 12 (2011) 2493–2537.

[13] J. B. Tenenbaum, V. d. Silva, J. C. Langford, A global geometric framework for nonlinear dimensionality reduction, science 290 (2000) 2319–2323.

[14] G. Sidorov, F. Velasquez, E. Stamatatos, A. Gelbukh, L. Chanona-Hernández, Syntactic n-grams as machine learning features for natural language processing, Expert Systems with Applications 41 (2014) 853–860.

[15] S. García, S. Ramírez-Gallego, J. Luengo, J. M. Benítez, F. Herrera, Big data preprocessing: methods and prospects, Big Data Analytics 1 (2016) 1–22.

[16] B. Charbuty, A. Abdulazeez, Classification based on decision tree algorithm for machine learning, Journal of Applied Science and Technology Trends 2 (2021) 20–28.

[17] K. P. Sinaga, M.-S. Yang, Unsupervised k-means

clustering algorithm, IEEE access 8 (2020) 80716–80727.

[18] F. Murtagh, P. Contreras, Algorithms for hierarchical clustering: an overview, ii, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 7 (2017) e1219.

[19] N. M. Abdulkareem, A. M. Abdulazeez, D. Q. Zeebaree, D. A. Hasan, Covid-19 world vaccination progress using machine learning classification algorithms, Qubahan Academic Journal 1 (2021) 100–105.

[20] G. Biau, E. Scornet, A random forest guided tour, Test 25 (2016) 197–227.

[21] M. Tanveer, T. Rajani, R. Rastogi, Y.-H. Shao, M. Ganaie, Comprehensive review on twin support vector machines, Annals of Operations Research (2022) 1–46.

[22] M. Azizi, S. Talatahari, A. H. Gandomi, Fire hawk optimizer: A novel metaheuristic algorithm, Artificial Intelligence Review 56 (2023) 287–363.

[23] N. E. H. B. Chaabene, A. Bouzeghoub, R. Guetari, H. H. B. Ghezala, Applying machine learning models for detecting and predicting militant terrorists behaviour in twitter, in: 2021 IEEE international conference on systems, man, and cybernetics (SMC), IEEE, 2021, pp. 309–314.