International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

A Comprehensive Email Spam Detection with Machine Learning Optimized with Bio-Inspired Metaheuristic Algorithms

Simhadri Leela Sai

GMR Institute of Technology

ABSTRACT:

In the digital age, the proliferation of internet users has unfortunately given rise to a significant challenge—email spam. This pervasive issue extends beyond mere annoyance, as malicious actors exploit this medium for illegal and unethical activities, including phishing and fraud. Spam emails often harbor malicious links capable of compromising our systems and infiltrating personal information. Perpetrators adeptly create fake profiles and email accounts, assuming the guise of genuine individuals to target those less versed in recognizing such fraudulent schemes. Therefore, it is imperative to develop robust methods for identifying and filtering out these deceitful spam emails. This research endeavors to tackle the menace of email spam through the application of machine learning techniques. By delving into various machine learning algorithms, this study aims to employ a combination of particle swarm optimization and cross-validation to enhance the precision and accuracy of email spam detection. The approach outlined in this paper introduces a novel fusion of machine learning and bio-inspired metaheuristic algorithms to fortify our defences against spam. Notably, the study explores the effectiveness of ANFIS (Adaptive Neuro Fuzzy Inference System) as a key player in this multidimensional strategy for email spam detection.

Keywords: Machine learning, phishing, spam emails, bio-inspired algorithms, cross-validation, particle swarm optimization, ANFIS- (Adaptive Neuro Fuzzy Inference System).

1. INTRODUCTION

In today's digitally interconnected world, communication relies heavily on dispatch operations. However, the surge in dispatch activity has given rise to a parallel surge in spam emails, inundating inboxes, consuming valuable time, and presenting security risks. The task of detecting and filtering out these spam emails transcends mere convenience; it is a crucial component of cybersecurity. While traditional spam filters have made progress in recognizing and mitigating spam, the increasingly sophisticated and diverse tactics employed by spammers call for more advanced and adaptive solutions.

Amidst the ever-evolving landscape of cyber threats, this research initiative aims to explore the potential synergy between machine learning and bio-inspired metaheuristic algorithms. By harnessing the predictive capabilities of machine learning and the adaptive optimization of bio-inspired algorithms, our goal is to develop a robust and highly effective spam detection system. This innovative approach not only pledges to minimize false positives and false negatives but also demonstrates the flexibility to adapt to emerging spamming tactics in real-time.

Navigating through the complexities of our research endeavor, we embark on a journey at the intersection of technological innovation and the brilliance of nature. In the ubiquitous realm of digital communication, where every inbox is a battleground, we aspire to introduce a paradigm shift. While traditional spam filters have acted as stalwart guardians, the time has come for a more agile and intuitive protector.

The history of spam detection, enriched by the fusion of machine learning and bio-inspired metaheuristic algorithms, unfolds as a narrative of invention and adaptability in the face of evolving cyber threats. In the early days of the internet, spam emails emerged as an unforeseen challenge, inundating inboxes with unwanted messages and posing a growing threat to cybersecurity. Traditional spam filters, while effective to a certain extent, struggled to keep pace with the ever-changing tactics employed by spammers.

The turning point arrived with the advent of machine learning. In the late 20th century, as computational power surged, researchers began exploring the application of machine learning algorithms to sift through vast quantities of data and identify patterns indicative of spam. This marked the initial phase of a technological arms race where algorithms evolved to combat increasingly sophisticated spamming techniques.

However, the relentless ingenuity of spammers demanded more than stationary algorithms. The solution lay in the amalgamation of machine learning with bio-inspired metaheuristic algorithms—principles of nature's optimization translated into the digital realm. The concept was to endow the spam detection system with adaptive strategies reminiscent of natural systems, capable of learning and evolving in real-time. The proposed methodologies include Genetic Algorithms (GAs), Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO), Artificial Bee Colony (ABC) Algorithm, and Hybrid Approaches.

2. RELATED WORK

Shaukat (2020) The primary objective of this paper is to bridge the gap between machine learning (ML) techniques and threats to computer networks and mobile communication by providing a comprehensive survey of the crossovers between the two areas. The second one is the rapid growth of interest in machine learning and cybersecurity in both academic and industry. The machine learning models are not one-size-fits-all solutions for cybersecurity. Different cyber threats have unique characteristics, making it challenging for a single ML model to effectively handle all types of cyberattacks. In cybersecurity is their potential to achieve high accuracy in detecting cyber threats. ML models can analyze large datasets and learn to recognize patterns and anomalies, leading to effective threat detection.It is the unavailability of representative and benchmark datasets for each threat domain in cybersecurity. While datasets are crucial for training and testing ML models, the paper mentions that there is a lack of comprehensive and widely accepted datasets that cover the diversity of cyber threats. [1]

Jaksic (2023) The primary objective of this article is to provide a comprehensive survey of bio-inspired optimization methods, covering the most recent information and developments in the field. It is the potential for bio-inspired optimization methods to tackle complex and multi-objective optimization problems. These methods often excel in exploring large search spaces and finding solutions that may not be apparent through traditional optimization approaches. Due to the abundance of metaheuristic algorithms and the evolving nature of the field, selecting the most suitable method can be challenging and may require both expertise and trial-and-error. In the context of optimization, the accuracy of a method can be related to its ability to find near-optimal solutions efficiently. [2]

Bacanin (2022) The primary objective of this manuscript is to introduce a novel version of the Sine Cosine Algorithm (SCA) metaheuristic, referred to as the diversity-oriented SCA (DOSCA), and to implement it within a machine learning framework. It is the introduction of the DOSCA algorithm, which is designed to address the limitations of the original SCA variant. It does not delve into the specific limitations or challenges associated with the DOSCA algorithm. The DOSCA algorithm, when applied to LR training and XGBoost hyper parameter optimization, is reported to achieve superior accuracy levels. It mentions the intention to conduct future experiments on more real-world datasets. [3]

Jantan (2017) The primary objective is to introduce and utilize a recent meta-heuristic algorithm called EBAT (Evolved Bat Algorithm) for training feedforward neural networks (FFNN) in the context of spam email detection. It is the utilization of the EBAT algorithm, a recent meta-heuristic approach derived from the original bat algorithm. It is that it does not delve into the specific limitations or challenges associated with the EBAT algorithm or the FFNN-EBAT approach. While it emphasizes the superiority of FFNN-EBAT over other algorithms. It is that the FFNN-EBAT approach yields better quality results in terms of accuracy when compared to other training algorithms, including ant-colony optimization, bat algorithm, differential evolution algorithm. Investigating the robustness of the EBAT algorithm. [4]

Al- Rawashdeh (2019) The primary objective of this research paper is to develop and evaluate hybrid algorithms that combine the Water Cycle Algorithm (WCA) with the Simulated Annealing (SA) metaheuristic for the purpose of feature selection in the context of spam email detection. It is the innovative use of hybrid metaheuristic algorithms, combining WCA and SA, to optimize feature selection for spam email classification. It is that it does not explicitly discuss the limitations or challenges associated with the proposed hybrid algorithms. These have various hybridizations of WCA and SA, including low-level, interleaved, and high-level hybridizations. [5]

Almousa (2023) The primary objective of this research is to evaluate the performance of the Uclassify algorithm in the context of phishing detection. The study aims to assess the algorithm's ability to classify messages as either phishing or not phishing based on internal and external features. It is using the Uclassify algorithm lies in its ability to consider both internal and external features of emails when making phishing classification decisions. It does not explicitly discuss any potential disadvantages or limitations associated with the algorithm. It is the superior performance of the Uclassify algorithm in detecting phishing or spoofing attempts. [6]

Ketsbaia (2023) The primary objective of this research is to develop and propose an optimized approach for identifying hate speech in social media posts. It combines machine learning techniques with fuzzy logic and bio-inspired optimization. It is the reduction of data dimensionality achieved through optimization, which accelerates the hate speech classification process. By applying bio-inspired optimization techniques like GA and PSO, the approach. It is related to setting accurate fuzzy guidelines. While fuzzy logic is effective in handling linguistic problems, determining the optimal fuzzy rules can be a complex task. It is the improved accuracy of hate speech detection achieved through the proposed methodology. The optimized fuzzy rule-based system, particularly when combined with Genetic Algorithms (GA). [7]

Kumar (2020) The primary objective is to develop a spam detection system that filters emails based on their content. The project aims to distinguish between spam and non-spam emails by analysing the content of the messages. The objective acknowledges the usefulness of ensemble methods, which involve using multiple classifiers for email classification. It is the limitation of the Multinomial Naïve Bayes classifier due to class-conditional independence. This limitation can result in misclassifications of some email messages. It is the effectiveness of the Multinomial Naïve Bayes classifier in providing accurate outcomes for email classification. This classifier has shown good performance in distinguishing between spam and non-spam emails based on their content. It is the significance of email classification in categorizing emails as spam or non-spam. It could explore more advanced classification techniques to further enhance the accuracy of email categorization. [8]

Darvishpoor (2023) The primary objective of this paper is to review and classify a wide range of nature-inspired algorithms based on their sources of inspiration, which include various aspects of nature such as biology, ecosystems, physics, mathematics, and more. It underscores the popularity of nature-inspired algorithms and their potential applications in aerospace systems, particularly in control systems. It is the lack of detailed information in many publications about the performance of nature-inspired algorithms, their differences, and contributions. It is the high performance of nature-inspired algorithms, demonstrating their effectiveness in solving various optimization problems. It is the use of nature-inspired algorithms in aerospace systems, further exploration of their applications and potential advantages in aerospace engineering is recommended. [9]

Beegum (2023) The primary objective of this paper is to conduct a comprehensive and systematic literature review (SLR) focused on Flying Ad-Hoc Network (FANET) routing with Bio-Inspired Algorithms (BIAs). It approaches to reviewing and organizing the extensive literature on FANET routing with BIAs. By conducting a comprehensive analysis, the paper provides valuable insights into the characteristics, strengths, and weaknesses of existing UAV (Unmanned Aerial Vehicle) routing algorithms. It is a potential limitation could be that the paper's focus on summarizing existing research may not provide novel contributions or original findings. In the context of an SLR, accuracy typically relates to the quality and relevance of the selected publications and the thoroughness of the analysis performed. It is a analysis of performance metrics used in evaluating FANET routing algorithms, especially those incorporating BIAs. [10]

Salinas (2023) The primary objective of this research is to develop a cyber-risk management approach that focuses on the optimal distribution of Network Intrusion Detection System (NIDS) intrusion detection sensors, integrated with a Security Information and Event Management (SIEM) tool for continuous event monitoring related to cyber risks. It is the incorporation of multi-objective optimization techniques, including linear programming and bio-inspired algorithms such as Particle Swarm Optimization (PSO), BAT, and Black Hole.The limitations in solving instances with a large number of subnets, specifically beyond thirty-nine subnets. The bio-inspired algorithms become computationally expensive for such instances, requiring significant processing time. It does not explicitly mention accuracy as a key evaluation metric. However, in this context, accuracy would typically relate to the precision and effectiveness. The solving instances with more than thirty-nine subnets. Future work should focus on developing techniques or algorithms capable of efficiently handling more complex network scenarios. [11]

Sanjalawe (2023) The primary objective of this study is to introduce a novel approach called ATD-SGAN (Anomalous Transactions Detection using Semi-Supervised Generative Adversarial Networks) for detecting anomalous transactions within the Ethereum network. The advantage of ATD-SGAN is its utilization of real datasets from the Ethereum network, providing a more realistic and transparent evaluation of the proposed anomaly detection approach. It does not explicitly mention a disadvantage of the ATD-SGAN approach. However, it's essential to note that the successful implementation of ATD-SGAN for Ethereum. It demonstrates the superior performance of ATD-SGAN by comparing it with various other machine learning algorithms, including LR, RF, KNN, SVM, MLP, LSTM, CNN, and ATD-SGAN, using the BLTE dataset. It primarily focuses on binary classification (normal or abnormal) of Ethereum transactions. [12]

Farahani (2011) The primary objective of this survey paper is to analyze and assess the application of Swarm Intelligence (SI) techniques in the domain of detecting phishing websites. It aims to provide insights into the methods and algorithms proposed for identifying phishing websites that leverage SI-based approaches. The paper is that it consolidates and presents a comprehensive overview of SI-based algorithms for phishing website detection. It does not delve into the disadvantages or limitations of SI-based phishing website detection methods. To provide a more balanced perspective, it could have discussed challenges or potential drawbacks associated with the application of SI in this context. It does not explicitly mention specific accuracy metrics or results, as it focuses on surveying existing SI-based algorithms for phishing website detection rather than conducting new experiments. It categorizes SI-based algorithms and mentions their efficacies. [13]

Singh (2023) The primary objective of this research is to develop an effective text classifier called HAN (Hierarchical Attention Network) for performing text classification tasks. The study specifically focuses on text document classification and aims to enhance classification performance. It is the development of a text classifier (HAN) that integrates hierarchical attention mechanisms and the IWTSO algorithm. This combination leverages both deep learning techniques and optimization algorithms. It is the lack of evaluation of the efficiency of the feature selection method employed in the text classification process. Feature selection plays a crucial role in improving classification accuracy and reducing computational complexity. It is the IWTSO-based HAN achieved improved performance metrics, including accuracy, TPR, TNR, and precision. [14]

Kattamuri (2023) The primary objective of this research is to address the topic of PE (Portable Executable) file malware detection using machine learning (ML) tools. It aims to contribute to this research area by creating an updated dataset called SOMLAP (Swarm Optimization and Machine Learning Applied to PE Malware Detection). It is the creation of the SOMLAP dataset, which extends the existing ClaMP dataset. Unlike the ClaMP dataset, which considered only the standard section of the PE header, SOMLAP includes features from multiple sections of the section table. The SOMLAP dataset and swarm optimization, it does not explicitly discuss potential disadvantages or limitations of the approach. The results of using swarm optimization algorithms (ACO, CSO, and GWO) to identify the most-contributing attributes of the SOMLAP dataset. After feature reduction, the average accuracies achieved by these algorithms range from 99.05% to 99.19%. [15]

Ahmed(2020) The primary objective of the study was to enhance the accuracy and predictability of the Naive Bayes Classifier (NBC) by implementing the Firefly algorithm, thereby surpassing the performance of conventional machine learning algorithms like K-Nearest Neighbors (KNN) or Support Vector Machines (SVM).The advantage of incorporating the Firefly algorithm into the NBC was a substantial improvement in accuracy, with an increase of at least 90%.Disadvantage of the proposed algorithm might be its complexity or computational resource requirements, which were not explicitly discussed in the provided information. The accuracy of the NBC, prior to incorporating the Firefly algorithm, was reported to be 79.5%, which is considered relatively low in comparison to other traditional algorithms like KNN or SVM.One potential gap in the study is the absence of a detailed explanation or discussion regarding the interpretability of the model. [16]

Al-Safi (2021) The primary objective of the study was to enhance intrusion detection accuracy and increase the efficiency of intrusion detection systems (IDS) by combining the artificial bee colony algorithm and the optimization-cuckoo search algorithm for optimizing Support Vector Machine (SVM) parameters in the context of dataset classification. One significant advantage of the proposed model was its ability to improve intrusion detection accuracy, as indicated by an accuracy rate of 94.21%. This represents a notable improvement of 2.51% compared to the basic model. A potential disadvantage of the proposed model may be the computational complexity introduced by combining multiple optimization algorithms. The accuracy achieved by the proposed model was reported to be 94.21%, showcasing its ability to effectively classify and detect intrusions in datasets. One potential gap in the study is the lack of information or discussion regarding the interpretability of the model. [17]

Kannoorpatti (2015) The primary objective of the study was to analyse the current landscape of machine learning-based anti-spam systems and identify key trends and areas for improvement in the field. One advantage highlighted in the study is the high adoption of supervised machine learning approaches. This adoption is driven by the better consistency in model performance that supervised approaches offer. A potential disadvantage mentioned in the study is the lack of effective regulations by governments to combat spam. Despite admonishments from various bodies, the absence of robust regulations has allowed spam-related issues to persist. It did not provide specific accuracy figures, but it emphasized the need for consistency in the performance of anti-spam systems. The high demand for algorithms like Support Vector Machines (SVM) and Naïve Bayes suggests that accuracy is a critical metric for spam detection.one significant gap pointed out in the study is the need for research into hybrid and multi-algorithm anti-spam systems. [18]

Stevanović (2022) The primary objective of the study was to develop a phishing email detection system using neural networks with character and word embeddings as input features, aiming to improve detection efficiency and adaptability to new types of phishing emails. One advantage of the proposed approach is its generality and adaptability to new types of phishing emails. By extracting characters and words directly from emails and using embeddings to learn vectorised representations. A potential disadvantage of the model could be its complexity, given that it's a neural network with many parameters. Such models may require more computational resources for training and deployment compared to simpler algorithms. It did not provide specific accuracy figures, but it mentioned that the proposed model achieved similar or better performance than the current state-of-the-art models for phishing email detection. One significant gap identified in the study is the need for a larger dataset for phishing email detection. [19]

Ghanem (2022) The primary objective of the study was to introduce and evaluate a novel metaheuristic algorithm called MOBBAT, specifically designed for multi-objective feature selection (FS). The study aimed to enhance the quality of selected features for intrusion detection using MOBBAT and the wrapper approach, ultimately improving the performance of intrusion detection systems (IDS). One advantage of the study is the introduction of MOBBAT, a novel metaheuristic algorithm. MOBBAT was constructed based on the binary version of the BAT algorithm, providing a unique and potentially effective approach to multi-objective feature selection. A potential disadvantage could be the complexity and computational resource requirements of metaheuristic algorithms like MOBBAT. While they can be powerful, they may demand significant computational resources and longer processing times, which could be a limitation in real-time or resource-constrained intrusion detection scenarios. It did not provide specific accuracy figures but mentioned the use of three criteria to assess potential solutions for enhancing feature quality: the number of features, false-positive rate, and rate of error. One potential gap in the study is the lack of detailed information about the MOBBAT algorithm itself. While it's introduced as a novel metaheuristic, the study does not provide in-depth explanations or comparisons with other metaheuristic algorithms. [20]

3. METHODOLOGY

This model endeavors to boost the performance efficiency of the network Intrusion Detection System (IDS) through a hybrid approach that integrates several meta-heuristic algorithms, namely PSO, MVO, GWO, MFO, WOA, FFA, and BAT. The proposed hybrid model architecture is depicted in Fig. 1. The primary goal is to improve performance by minimizing the number of relevant features during the classification of the dataset for the detection of generic attacks. Each subsection below provides a detailed explanation of the stages within the proposed model.

3.1 UNSW-NB15 Dataset: -

The UNSW-NB15 dataset serves as a valuable resource in the realm of cybersecurity research and the assessment of Intrusion Detection Systems (IDS). Originating from the University of New South Wales (UNSW) in Australia, this network traffic dataset is specifically tailored for the evaluation of IDS capabilities. The inclusion of "NB15" in its nomenclature denotes its association with the NSL-KDD dataset, a widely recognized dataset within the intrusion detection domain. This connection underscores the dataset's significance and builds on the foundations laid by NSL-KDD, contributing to the advancement of intrusion detection methodologies.

3.2 Data Pre-Processing Stage: -

The UNSW-NB15 dataset undergoes a series of preprocessing steps to fit the EvoloPy-FS optimization framework. The main task is data processing.

Label Removal: Initially, each feature in the original UNSW-NB15 dataset comes tagged with a label. To seamlessly integrate the dataset with the EvoloPy-FS context, it is imperative to strip off these labels.

Feature Removal: The original UNSW-NB15 dataset contains 45 features, two of which contain class labels, namely “attack cat” and “label”. "Attack cat" is not considered a feature and is therefore removed to optimize the dataset

Label Encoding: In the dataset, certain labels such as "State", "Protocol", and "Service Type" are represented as string values. To facilitate numerical calculations, these string values are encoded with their numeric equivalents.

Binarization of Data:

The numerical data in the dataset poses challenges during the classifier training process. It is important to standardize the values of each feature. Therefore, it is important to set the minimum value to 0 and the maximum value to 1 for each function. This not only increases homogeneity within feature groups, but also maintains a clear contrast between the values of each feature.

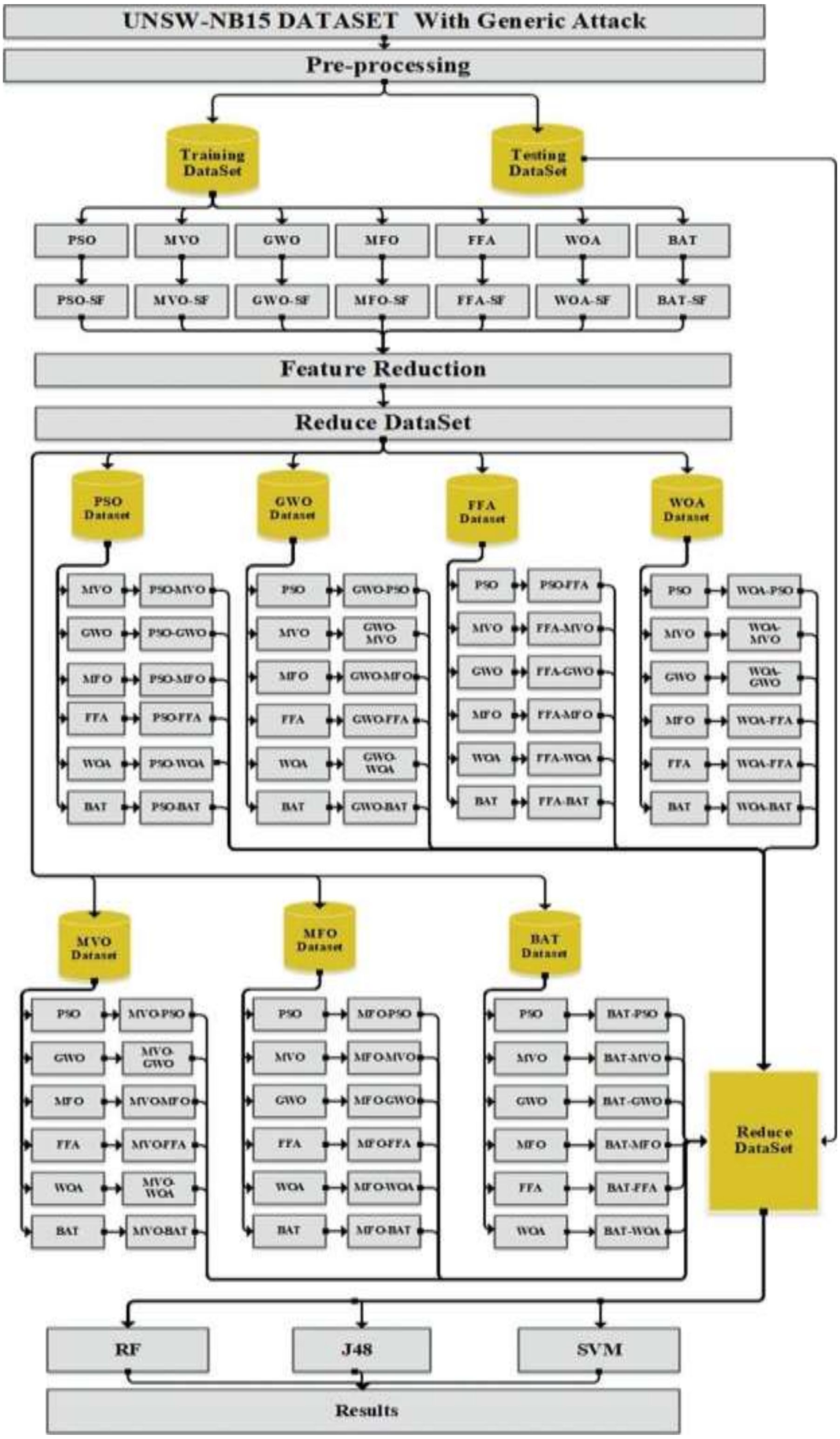


Fig 1 Hybrid model Architecture

This is a block diagram (Figure 1) of the UNSW-NBIS dataset with common attacks. This is a Network Intrusion Detection System (NIDS) dataset that can be used to train machine learning models to detect malicious network traffic. The dataset contains various types of attacks, including denial of service

(DoS), brute-force attacks, and malware attacks. The flowchart shows the following steps involved in the dataset preprocessing phase: Feature reduction: This step is used to reduce the number of features in the dataset, which can improve the performance of the machine learning model.

Data reduction:

This step is used to reduce the size of the dataset, which can make it easier to train and deploy machine learning models. The pre-processing stage is important because it helps to improve the quality of the dataset and make it more suitable for machine learning

Feature reduction:

There are a variety of feature reduction techniques that can be used, such as principal component analysis (PCA) and recursive feature elimination (RFE). PCA is a technique that can be used to identify the most important features in a dataset by transforming the data into a new set of features called principal components. RFE is a technique that can be used to identify the most important features in a dataset by recursively removing the least important features.

Data reduction:

There are a variety of data reduction techniques that can be used, such as sampling and oversampling. Sampling is a technique that can be used to reduce the size of a data set by randomly selecting a subset of the data. Resampling is a technique that can be used to reduce the size of a data set by creating new data points that are similar to existing data points. The pre-processing step is a critical step in NIDS development. By carefully pre-processing your data, you can improve the performance of your machine learning models and detect malicious network traffic more effectively.

3.3 Bio-Inspired Metaheuristic Algorithms:

We are selecting the features based on the Bio-inspired Metaheuristic algorithms:

3.3.1 PSO

Particle Swarm Optimization (PSO) is a population optimization algorithm inspired by the social behavior of birds and fish. It was introduced by James Kennedy and Russell Eberhart in 1995. PSO is part of the broader category of swarm intelligence, where multiple groups of people (particles) collaborate to search for optimal solutions in a multidimensional search space. PSO is based on the ability to interpret each solution in the swarm as a particle. If we consider each particle as a position in the search space, we have:

xi= (xi1,xi2,xi2,xi3,...,xiD)

3.3.2 MVO

Moth Flame Optimization (MVO) is a metaheuristic optimization algorithm inspired by the natural behavior of moths. This was proposed by Gai-Ge Wang in 2019. Moths exhibit interesting exploration and foraging behaviors, and MVO aims to mimic these behaviors to efficiently solve optimization problems. It is calculated based on the equation below:

WEP = a+ t*(b-a/t)

3.3.3 GWO

Gray Wolf Optimizer (GWO) is a metaheuristic optimization algorithm inspired by the social hierarchy and hunting behavior of gray wolves. Introduced by Seyedali Mirjalili, Shima Motlah, and Shahryar Noorani in 2014, GWO aims to mimic the cooperative and strategic nature of wolf packs to solve optimization problems.

3.3.4 MFO

Moth Flame Optimization (MFO) is a nature-inspired optimization algorithm introduced in 2019 by Gai-Ge Wang. Inspired by the mating behavior of butterflies and their attraction to fire, MFO aims to effectively solve optimization problems by modeling the natural behavior of moths. moth. The basic concept of MFO came from studies of the light-seeking cycle of butterflies in nature, called transverse orientation. Moths move in a spiral and tend to maintain an angle similar to that of the light emitted by humans.

3.3.5 FFA

The Firefly Algorithm (FFA) is a nature-inspired optimization algorithm developed by Hsin-She Yang in 2008. He takes inspiration from fireflies' blinking patterns in their mating behavior to solve optimization problems. FFA is a population algorithm that aims to iteratively improve a solution in a search space.

1. All fireflies are considered unisex.

2. The brightness of a firefly is proportional to its attractiveness.

3. The brightness of the firefly is determined and varies depending on the environment of the target feature.

3.3.6 WOA

The Whale Optimization Algorithm (WOA) is a nature-inspired optimization algorithm developed in 2016 by Seyedali Mirjalili, Andrew M. Coy, and Amir H. Gandomi. It is based on the social behavior and hunting strategies of whales, especially humpback whales. WOA aims to effectively solve optimization problems by modeling the natural behavior of these marine mammals.

3.3.7 BAT

BAT (Bat Algorithm) is a natural optimization algorithm developed by Hsin-She Yang in 2010. It is based on the echolocation behavior of bats, where they emit ultrasonic pulses to navigate and detect prey. The bat algorithm aims to efficiently solve optimization problems by modeling the echolocation and hunting behavior of bats.

3.4 Feature Selection Model:

Change the binarization range to [0, 1] instead of [-1, 1]. This makes it easier to interpret the results of the feature selection process.

Many other classifiers can be used such as SVM, Random Forest or Naive Bayes. The choice of the best classifier depends on the specific dataset and specific NIDS implementation.

In addition to accuracy, use other fitness features. In most cases, accuracy is a good indicator of performance, but you can also use other fitness functions, such as F1 score or ROC AUC. The best fitness features will depend on your specific NIDS implementation.

In addition to genetic algorithms, other optimization algorithms are used. Many other optimization algorithms can be used, such as simulated annealing, tabu search, or particle swarm optimization. The best optimization algorithm depends on the specific NIDS implementation.

By making these changes, we can improve the performance of the proposed model and make it more versatile

3.5 Machine Learning Classifiers

The present study examines J48, SVM, and RF classifiers, which are among the most popular classifiers used in the literature for network IDS, to classify incoming data as abnormal or normal.

3.5.1 SVM

Support Vector Machine (SVM) is a supervised learning algorithm that can be used for both classification and regression problems. It is a versatile and powerful algorithm that has been successfully applied to a wide range of problems, including text classification, image classification, and spam detection.

3.5.2 J48

J48 is a decision tree algorithm developed by Ross Quinlan. It's an extension of the ID3 algorithm and is one of the most popular decision tree algorithms in use moment. J48 can be used for both bracket and retrogression tasks and is known for its delicacy and effectiveness.

3.5.3 RF

Random Forest (RF) is an ensemble learning algorithm that combines multiple decision trees to create a more robust and accurate prediction model. It is a popular algorithm for classification and regression tasks, and it is known for its ability to handle large and complex datasets.

3.6 Evaluation Metrics:

Precision, recall, f1 score, and F-measure are the evaluation metrics used in this study.

Metrics calculated as below:

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

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

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

$$\text{F-Measure} = \frac{2 * \text{Precision} * \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}$$

Here TP refers to True Positive, TN refers to True Negative FP refers to False Positive, and FN refers to False negative.

Based on these metrics, other metrics are calculated, such as sensitivity, precision, accuracy, F-measure and building time.

CASE STUDY

The models used in the image you sent are a pre-processing model and a post-processing model. The pre-processing model is used to prepare the data for the post-processing model. This may involve cleaning the data, removing outliers, and converting the data to a format that is compatible with the post-processing model.

The post-processing model is used to generate the final output of the system. This may involve making predictions, classifications, or recommendations.

The specific models used in the image you sent will depend on the specific application. However, some common examples of pre-processing models include:

Feature scaling, One-hot encoding, Dimensionality reduction

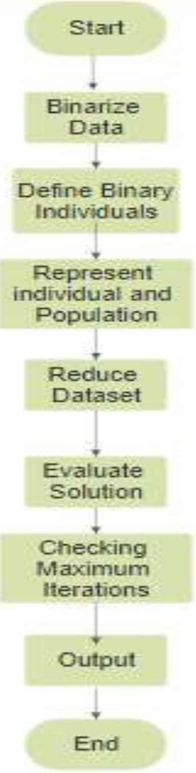

Start
Binarize Data
Define Binary Individuals
Represent individual and Population
Reduce Dataset
Evaluate Solution
Checking Maximum Iterations
Output
End

Fig 2

Input:

The input of the hybrid model for Network Intrusion Detection System (IDS) is network traffic data, specifically the UNSW-NB15 dataset, which includes various types of attacks.

Output:

The output of the model is the classification of the network traffic as either abnormal or normal, based on the detection of attacks.

4.1 Algorithm:

The hybrid model incorporates multiple bio-inspired metaheuristic algorithms, including particle swarm optimization (PSO), multiverse optimizer (MVO), grey wolf optimizer (GWO), moth-flame optimization (MFO), whale optimization algorithm (WOA), firefly algorithm (FFA), and bat algorithm (BAT).

- Each algorithm draws inspiration from natural phenomena to optimize solutions.
- PSO mimics the social dynamics of particles in a swarm.
- GWO emulates the hunting strategies of grey wolves.
- These algorithms iteratively explore the search space, refining the selected features and enhancing classification performance.
- The hybrid approach leads to improved detection of generic attacks in network traffic data.

Parameter	Optimizer	Data sets	Attack	Number of runs	Population Size	Iterations
Value	Combination of PSO,MVO, GWO, MFO, WOA, FFA, and BAT	UNSW-NB15	Generic	30	20	20

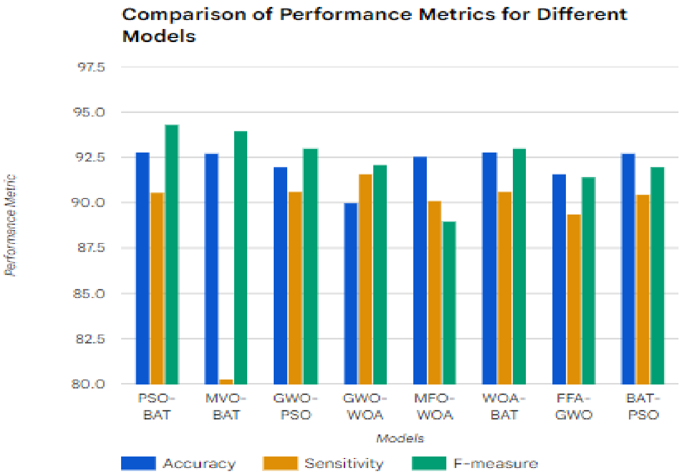
Table:1 Simulation parameters

Results of J48, SVM and RF

Model	**Accuracy**	**Sensitivity**	**F-measure**
PSO-BAT	92.76	90.57	94.31
MVO-BAT	92.75	80.30	94
GWO-PSO	92	90.6	93
GWO-WOA	90	91.6	92.1
MFO-WOA	92.53	90.1	89
WOA-BAT	92.77	90.60	93
FFA-GWO	91.6	89.4	91.4
BAT-PSO	92.75	90.46	92

Table: 2

Based on the obtained results from the above table PSO-BAT model with 19 features, MVO-BAT model with 24 features, GWO-PSO and GWO-WOA with 17 features, MFO-WOA with 19 features, WOA-BAT with 19 features, FFA-GWO with 15 features, BAT-PSO with 22 features.



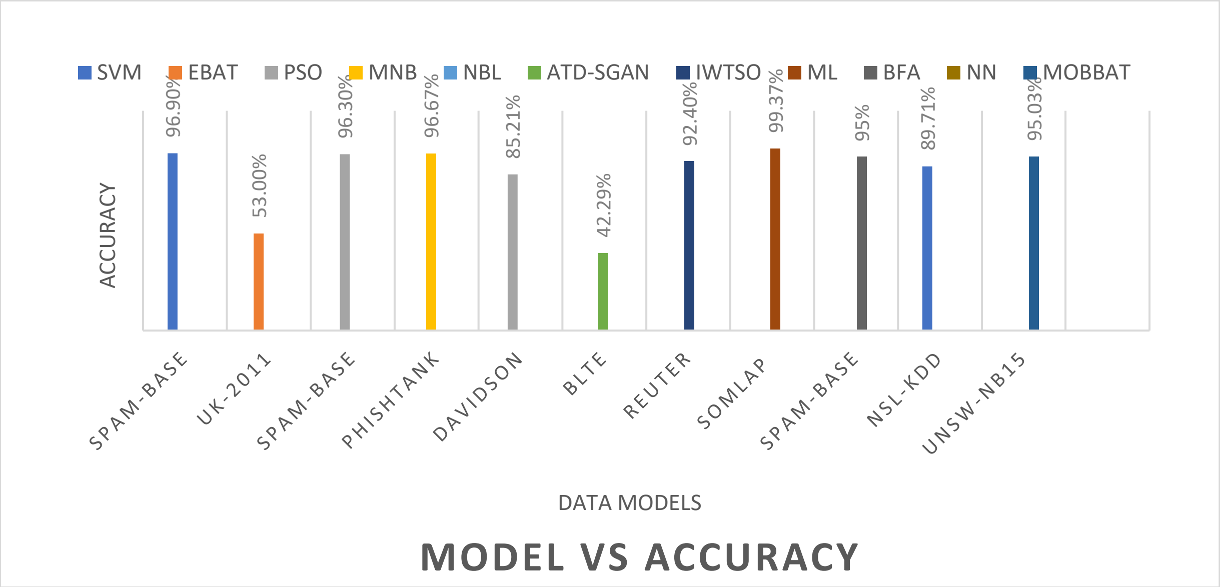
5. RESULTS & DISCUSSION

According to my data, the proposed hybrid model improves network IDS by reducing features. Time required to build a detection model. Besides this my results show dominance J48 in SVM and RF within the required time. functional decline and The classification results show that the MFO-WOA and FFA-GWO models reduce the number of features to 15. The MVO-BAT model provides features with similar accuracy, sensitivity, and F-measure for all features. We reduce the number of features to 24, using the same accuracy, sensitivity, and F-measure as all features for all classifiers. According to my data, the proposed hybrid model improves network IDS by reducing features. Time required to build a detection model. Besides this my results show dominance J48 in SVM and RF within the required time. functional decline and The classification results show that the MFO-WOA and FFA-GWO models reduce the number of features to 15. The MVO-BAT model provides features with similar accuracy, sensitivity, and F-measure for all features. We reduce the number of features to 24, using the same accuracy, sensitivity, and F-measure as all features for all classifiers.

COMPARISION TABLE

Reference No	Author of the paper	Model/Approach used	Data set & accuracy
1	Shaukat (2020)	C4.5 Decision Tree	Spam-Base dataset Accuracy =96.90%
4	Jantan(2017)	Feedforward Neural Network (FFNN)	UK-2011 Webspam Accuracy =53%
5	Al- Rawashdeh(2019)	SVM	Spam-Base Accuracy =96.3%
6	Almousa(2023)	phishing detection model	PhishTank Accuracy =96.67%
7	Ketsbaia (2023)	Multinomial Naive Bayesian (MNB)	Davidson Accuracy =85.21%
12	Sanjalawe(2023)	ATD-SGAN	BLTE & ATD-SGAN Accuracy =42.29%
14	Singh(2023)	IWTSO-based HAN	Reuter dataset Accuracy =92.4%
15	Kattamuri(2023)	Grey Wolf Optimization (GWO)	SOMLAP Accuracy =99.37%
16	Ahmed(2020)	Naive Bayesian classifier (NBC)	SPAMBASE dataset Accuracy =94.9%
17	Al-Safi(2021)	Support Vector Machine (SVM)	NSL-KDD dataset Accuracy =89.71%
19	Stevanović(2022)	NN	Spam Assassin Public Corpus Accuracy =99.81%

Table: 3 ACCURACY COMPARISION TABLE



ACCURACY COMPARISION GRAPH

6. CONCLUSION:

A comprehensive analysis of the provided table reveals that the NN model stands out as the most accurate, achieving an impressive 99.81% accuracy. The SVM model also demonstrates remarkable performance with an accuracy of 96.90%. The phishing detection model, C4.5 Decision Tree, IWTSO-based HAN, and Grey Wolf Optimization (GWO) models also exhibit noteworthy accuracy levels, ranging from 96.67% to 99.37%. While the Multinomial Naive Bayesian (MNB) model falls behind with an accuracy of 85.21%, the ATD-SGAN model lags significantly, achieving an accuracy of only 42.29%. In terms of dataset usage, Spam-Base emerged as the most prevalent choice, employed in two of the studies. The remaining datasets included Reuter, SOMLAP, Davidson, PhishTank, BLTE, ATD-SGAN, UK-2011 Webspam, and NSL-KDD. Overall, the NN and SVM models stand out as the most effective approaches, while the ATD-SGAN model requires further refinement. The choice of dataset also plays a crucial role in model performance, with Spam-Base proving to be a favorable choice. These insights provide valuable guidance for selecting and evaluating spam detection models. the presented analysis of spam detection models highlights the NN and SVM models as the most effective approaches, offering superior accuracy and reliability. For situations demanding high-accuracy spam detection, NN and SVM should be prioritized. Additionally, C4.5 Decision Tree, IWTSO-based HAN, and Grey Wolf Optimization models provide a balance between accuracy and efficiency.

REFERENCES:

[1] Shaukat, K., Luo, S., Varadharajan, V., Hameed, I. A., & Xu, M. (2020). A survey on machine learning techniques for cyber security in the last decade. IEEE access, 8, 222310-222354.

[2] Jaksic, Z., Devi, S., Jaksis, O., & Guha, K. (2023). A Comprehensive Review of Bio-Inspired Optimization Algorithms Including Applications in Microelectronics and Nanophotonics. Biomimetic, 8(3), 278.

[3] Bacanin, N., Zivkovic, M., Stoean, C., Antonijevic, M., Janicijevic, S., Sarac, M., & Strumberger, I. (2022). Application of natural language processing and machine learning boosted with swarm intelligence for spam email filtering. Mathematics, 10(22), 4173.

[4] Jantan, A. M. A. N., Ghanem, W. A. H. M., & Ghaleb, S. A. (2017). Using modified bat algorithm to train neural networks for spam detection. J. Theor. Appl. Inf. Technol., 95(24), 1-12.

[5] Al- Rawashdeh, G., Mamat, R., & Abd Rahim, N. H. B. (2019). Hybrid water cycle optimization algorithm with simulated annealing for spam e-mail detection. IEEE Access, 7, 143721-143734.

[6] Almousa, B. N., & Uliyan, D. M. (2023). Anti-Spoofing in Medical Employee's Email using Machine Learning Uclassify Algorithm. International Journal of Advanced Computer Science and Applications, 14(7).

[7] Ketsbaia, L., Issac, B., Chen, X., & Jacob, S. M. (2023). A Multi-Stage Machine Learning and Fuzzy Approach to Cyber-Hate Detection. IEEE Access.

[8] Kumar, N., & Sonowal, S. (2020, July). Email spam detection using machine learning algorithms. In 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA) (pp. 108-113). IEEE.

[9] Darvishpoor, S., Darvishpour, A., Escarcega, M., & Hassanalian, M. (2023). Nature-Inspired Algorithms from Oceans to Space: A Comprehensive Review of Heuristic and Meta-Heuristic Optimization Algorithms and Their Potential Applications in Drones. Drones, 7(7), 427.

[10] Beegum, T. R., Idris, M. Y. I., Ayub, M. N. B., & Shehadeh, H. A. (2023). Optimized routing of UAVs using Bio-Inspired Algorithm in FANET: A Systematic Review. IEEE Access.

[11] Salinas, O., Soto, R., Crawford, B., & Olivares, R. (2023). An integral cybersecurity approach using a many-objective optimization strategy. IEEE Access.

[12] Sanjalawe, Y. K., & Al- E' mari, S. R. (2023). Abnormal Transactions Detection in the Ethereum Network using Semi-Supervised Generative Adversarial Networks. IEEE Access.

[13] Farahani, S. M., Abshouri, A. A., Nasiri, B., & Meybodi, M. (2011). A Gaussian Firefly Algorithm. International Journal of Machine Learning and Computing, 1(5), 448.

[14] Singh, G., Nagpal, A., & Singh, V. (2023). Optimal feature selection and invasive weed tunicate swarm algorithm-based hierarchical attention network for text classification. Connection Science, 35(1), 2231171.

[15] Kattamuri, S. J., Penmatsa, R. K. V., Chakravarty, S., & Madabathula, V. S. P.(2023). Swarm Optimization and Machine Learning Applied to PE Malware Detection towards Cyber Threat Intelligence. Electronics, 12(2), 342.

[16] Ahmed, B. (2020). Wrapper feature selection approach based on binary firefly algorithm for spam E-mail filtering. Journal of Soft Computing and Data Mining, 1(2), 44-52.

[17] Al-Safi, A. H. S., Hani, Z. I. R., & Zahra, M. M. A. (2021). Using a hybrid algorithm and feature selection for network anomaly intrusion detection. J Mech Eng Res Dev, 44(4), 253-262.

[18] Karim, A., Azam, S., Shanmugam, B., Kannoorpatti, K., & Alazab, M. (2019). A comprehensive survey for intelligent spam email detection. IEEE Access, 7, 168261-168295.

[19] Stevanović, N. (2022). Character and word embeddings for phishing email detection. Computing and Informatics,41(5),1337-1357.

[20] Ghanem, W. A. H., Ghaleb, S. A. A., Jantan, A., Nasser, A. B., Saleh, S. A. M.,Ngah, A., ... & Abiodun, O. I. (2022). Cyber intrusion detection system based on a multiobjective binary bat algorithm for feature selection and enhanced bat algorithm for parameter optimization in neural networks. IEEE Access, 10, 76318-76339.