Detecting Spam Emails Using Machine Learning and Harris Hawks's Optimizer (HHO) Algorithm

B. RAJU[1], G. FATHIMA[2]
[1] *M.Sc., Department of Computer Science and Engineering, Dr. MGR Educational and Research Institute, Chennai, India*
[2] *Faculty Central of Excellence in Digital Forensics, Dr. MGR Educational and Research Institute, Chennai, India*

Abstract- ***Email spamming is an important issue in recent years. The number of spam emails increases as the number of internet users increases. Everyone is using technology for illegal and immoral acts like robbery and phishing. Spam phishing will make us mentally disturbed. As a result of it, the number of spam in our inbox will increase. It will affect our social life and make our life hell. One thing is sure we must control the spread of spam and use machine learning methods to trace the fraud spoofers. This paper discusses a new method for spam classification that combines a new metaheuristic technique called Harris Hawks Optimizer (HHO) and a machine learning technique called XG-Boost By using three sigmoid activation function and Logistic regression public database, the hybrid HHO-XG-boost models are proposed for spam classification. The Harris Hawks Optimization (HHO) algorithm is a new metaheuristic algorithm motivated by a natural cooperative hunting phenomenon of Harris's hawks and the HHO model that allows for an adaptive consideration of spatial and under-spatial shock search among neighbor spam-capturing datasets. From the experimental results, it is marked from figure tables that the proposed model is more accurate and attains better performance than others like noise removal. From the experimental procedures, the results are more accurate, novice, and fast than other antispam emails. From the experimental analysis, the proposed method's accuracy is very good since it reduced the spam drastically.***

Index Terms- ***Machine-learning algorithm (XG-boost), Detecting spam email, HHO Algorithm, spam email***

I. INTRODUCTION

In the era of information technology, information exchange has become much easier and faster. Many platforms allow users to share information from anywhere in the world. Of all information exchange media, email is the easiest, cheapest, and fastest way to share information around the world. However, due to its simplicity, email is vulnerable to various types of attacks, the most common and dangerous of which is spam [1]. No one wants to receive emails that are not relevant to their interests because they waste the recipient's time and resources. Additionally, these emails may contain malicious content in the form of attachments or URLs, which can lead to security compromises in the host system [2]. Spam is an unsolicited message or email sent by an attacker to a large number of recipients via email or other information exchange media [3]. Therefore, high demands are placed on the security of e-mail systems. Attackers typically use this technique to lure users to online services. It sends spam emails with "multi-file" extension attachments and packed URLs to redirect users to malicious spam websites and ultimately steal data, financial fraud, or personal information. It can lead to theft [4, 5]. Many email providers allow users to create keyword-based rules that automatically filter emails. Still, this approach is difficult and not very useful because users don't want to customize their emails. This allows spammers to attack your email account. The Internet of Things (IoT) has become a part of modern life in recent decades and is experiencing rapid growth. With the advent of IoT, spam problems are on the rise. Researchers have proposed various spam detection methods to detect and filter spam and spammers. These approaches have limitations and drawbacks. The rise of the Internet and global communication has led to a significant increase in spam emails [6]. The Internet is used to generate spam emails from anywhere in the world by disguising the identity of the attacker. Although there are many anti-spam tools and techniques, spam rates are still very high. The most dangerous spam

emails are malicious emails that contain links to malicious websites that can damage the victim's data. Spam emails can also use up your server's memory and capacity, causing slow response times. To accurately detect spam email and avoid the growing email spam problem, companies are carefully considering the anti-spam tools available in their environments. Well-known email identification and analysis mechanisms for spam detection include whitelisting/blacklisting [7], email header analysis, and keyword checking.

II. REVIEW OF LITERATURE

Sai Charan Lanka, Kodali Pujita, ET AL.,2024 had proposed Optimization of Naïve Bayes Classifier for Spam E-Mail Detection. Email is a popular and official communication platform and is widely used for its convenient and efficient way of facilitating the exchange of information between individuals or organizations and allowing any kind of information to be sent and received instantly from anywhere in the world. It is a communication method. However, with the rapid increase in email usage, spammers are using this platform to commit fraud via email, known as spam email. Spam emails can be detected and identified using various approaches. Among these approaches, machine learning is widely used. In machine learning, Naive Bayes classifiers have the highest accuracy due to their "low false positive rate." Naive Bayes Classifier provides the highest accuracy among all other machine learning models, but it can also be optimized for higher accuracy.

Mohammad Tubishat 1, Feras Al-Obeidat ET AL.,2023 had proposed An Improved Dandelion Optimizer Algorithm for Spam Detection. Next Generation Email Filtering System Spam email has become a widespread problem in recent years as Internet users increasingly receive unsolicited or fake email. To address this issue, automatic spam detection methods have been proposed that aim to classify emails into spam and non-spam categories. Machine learning techniques have been used for this task with great success.

Ala' m. Al-zoubi 1, antonio m. Mora et al.,2023 had proposed A Multilingual Spam Reviews Detection Based on Pre-Trained Word Embedding and Weighted Swarm Support Vector Machines. Online reviews are important information that customers look for when deciding to purchase a product or service. Additionally, organizations benefit from these reviews as important feedback on their products and services. Such information needed to be reliable, especially during the COVID-19 pandemic, when quarantines and shelter-in-place orders led to a significant increase in online reviews. Not only did the number of reviews increase to 4,444, but so too did the number of situations and preferences during the pandemic. Therefore, spam checkers are considering these changes and improving their deception techniques. Spam reviews typically consist of misleading, fake, or fraudulent reviews that are intended to trick customers into making money or harm other competitors. Therefore, in this study, we introduce Weighted Support Vector Machine (WSVM) and Harris Hawks Optimization (HHO) for spam review detection. HHO acts as an algorithm to optimize hyperparameters and feature weights.

Ashraf S. Mashaleh , Noor Farizah Binti Ibrahim et al.,2022 had proposed Detecting Spam Email with Machine Learning Optimized with Harris Hawks optimizer (HHO) Algorithm. Inspired by the hunting behavior of these birds, the Harris Hawks optimization algorithm opens a new perspective for solving complex optimization problems, in this case, spam email identification. By leveraging surprise jump tracking techniques observed in nature, the algorithm dynamically adjusts its search strategy to efficiently navigate high-dimensional data such as Spam-base datasets. With a spam detection accuracy of 94.3%, this new technology offers a promising solution to the growing challenge of combating email spam and protecting users from phishing and cyber threats.

Hidayet Takci, Fatema Nusrat,2022 had proposed Highly Accurate Spam Detection with the Help of Feature Selection and Data Transformation. While email is becoming more popular, the amount of spam is increasing rapidly. This situation has created a need to filter spam emails. To date, many knowledge-based, learning-based, and clustering-based methods have been developed to filter spam emails. This study targeted machine learning-based spam detection, and C4.5, ID3, RndTree, C-Support Vector Classification (C-SVC), and Naive Bayes algorithms were used to detect email spam detection. Additionally, feature selection and data transformation methods were used to increase the success rate of spam detection. Experiments were

conducted on the UC Irvine Machine Learning Repository (UCI) Spam-base dataset, and the results were compared in terms of accuracy, receiver operating characteristic (ROC) analysis, and classification speed. According to the accuracy comparison, the C-SVC algorithm showed the highest accuracy of 93.13%.

Na Song & Yunpeng Ma 2022 had proposed Harris Hawks optimization based on global cross-variation and tent mapping. To solve the problem of slow convergence due to the uniform selection position update formula in the exploration stage of the base HHO and trapping into local optimization due to lack of population richness in the later stages of the algorithm, the Harris-Hawks optimal In this article, we propose global cross-variation and tent mapping (CRTHHO). First, tent assignment is introduced in the exploration stage to optimize the random parameter q and accelerate the convergence in the initial stage. Second, a crossover mutation operator is introduced to cross over and mutate the global optimal position in each iteration process. A greedy algorithm from falling into a local optimum by skipping the optimal solution and improves the convergence accuracy of the algorithm

Simran Gibson , Biju Issac ET AL.,2020 had proposed Detecting Spam Email With Machine Learning Optimized With Bio-Inspired Metaheuristic Algorithms. Email has simplified communication for many organizations and individuals. This method is exploited by spammers to make a profit by sending unsolicited emails. The purpose of this article is to introduce how to detect spam emails using machine learning algorithms optimized in a bio-inspired manner. To optimize the performance of the classifier, biologically derived algorithms such as particle swarm optimization and genetic algorithms are implemented. Multinomial Naive Bayes with Genetic Algorithm showed the best overall performance. We also discuss a comparison of the results with other machine learning models and biologically inspired models to uncover the most appropriate model.

III. RESEARCH METHODOLOGY

This research paper suggests combining the best of both XGBOOST and HHO classifiers to improve our spam detection accuracy. The idea for this study was inspired by previous research comparing different hybrids of the Harris Hawks Optimizer with existing algorithms like XGB knowing very well what each model can do; hence selecting features is crucial. The main goal here is way better performance in terms of speed and precision when determining whether a message falls under junk or legitimate email therein thereby significantly increasing its' usefulness in society at large. In order to visualize entire set-up please refer to figure 1.

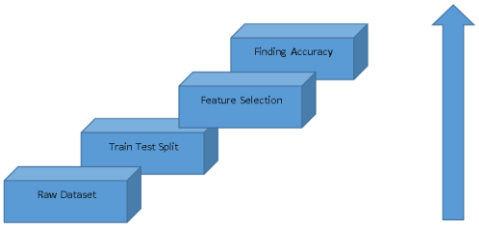

Fig. 1: architecture diagram

IV. DATA COLLECTION

The Spam-base dataset is one of the most popular datasets for machine learning and data mining research, especially in email spam detection. It has been sourced from the UCI Machine Learning Repository which is well-respected for providing a wide range of ML datasets. The Spam-base dataset holds email information characterized into two categories: spam and non-spam. In this dataset, there are 4,601 examples with 58 attributes. These attributes encompass different statistical measures taken from the body of an email like the frequency of specific words and characters, lengths of capital letter sequences as well as other important details. It is meant to supply a comprehensive set of data that can be used to train and test machine-learning models for recognizing email spam. Consequently, researchers apply it to develop algorithms and techniques to detect spam emails; and evaluate them using the same spam e-mail dataset collected earlier thereby making it a relevant resource for studying applications of machine learning in email filtering

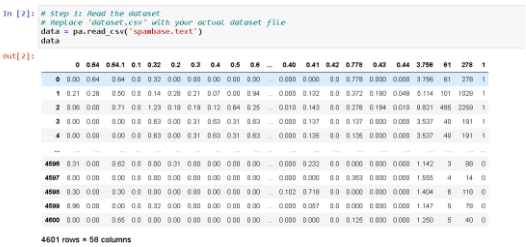

Fig. 2: Dataset Collection

V. TRAIN TEST SPLIT

It is also important to separate the gathered data into training and testing sets after data is collected; this is the second phase of the machine learning workflow process. Most joke about a train-test split to build and evaluate a predictive model. This is through the training set which is utilized to train the model and enable it to identify common structures in the dataset. It usually includes 70 – 80% of the overall dataset and helps the model to perform better in learning outcomes. Cross-validation performs a complete run of the model on the testing set to find the accuracy, precision, recall, and other assessment measures.

Train-test split can help to build reliable machine learning models that are not only accurate but also applicable to future sets of data in a particular environment to facilitate online predictions of a specific dataset.

VI. FEATURE SELECTION

Once we have split our training set into further testing sets, then eventually come to a point where we can determine the best subset for our model by simply analyzing certain chosen characteristic sets, which includes choice-based methods such as Forward, Backward, or Hybrid approaches which eliminate or add some attributes iteratively till there is a final solution. An example might be selecting two subsets with different numbers of dimensions available at each subset respectively; first selecting attributes which contribute most variance among all using p-values obtained from ANOVA then finally considering small features sets simultaneously fitting into some common model structure."

```
In [8]: # Step 3: Feature Selection Using HHO and XGBOOST
        def fitness(solution, X_train, y_train, X_test, y_test):
            # Your fitness function evaluation using k-NN here
            sf = np.where(solution == 1)[0]
            selected_features = X_train.iloc[:, sf]
            selected_features_test = X_test.iloc[:, sf]

            XGB_classifier = XGBClassifier()
            XGB_classifier.fit(selected_features, y_train)
            accuracy = XGB_classifier.score(selected_features_test, y_test)

            return accuracy

        def HHO_Feature_Selection(X_train, y_train, X_test, y_test, lb, ub, dim, SearchAgents_no, Max_iter):
            # Initialize the population of solutions (hawks)
            population = np.random.randint(2, size=(SearchAgents_no, dim))  # Randomly initialize solutions

            # Initialize convergence curve to track best fitness over iterations
            convergence_curve = np.zeros(Max_iter)

            # Loop through iterations
            for t in range(Max_iter):
                # Evaluate fitness for each solution in the population
                fitness_values = np.array([fitness(solution, X_train, y_train, X_test, y_test) for solution in population])

                # Find the index of the best solution
                best_index = np.argmax(fitness_values)
                best_fitness = fitness_values[best_index]
                best_solution = population[best_index]

                # Exploration and exploitation steps for each solution
                for i in range(SearchAgents_no):
                    # Check if the new solution is better and update
                    new_fitness = fitness(population[i], X_train, y_train, X_test, y_test)
                    if new_fitness > fitness_values[i]:
```

Fig. 3: Feature Selection using HHO and XG-Boost

VII. ALGORITHM

In our work, we focus on investigating boosting the performance of XG-Boost with HHO as a meta-heuristic search method. The XG-Boost is one of the best gradient-boosting frameworks that emphasizes prediction quality and efficiency. However, it is very sensitive to the hyperparameters used in the optimization process. To overcome this issue, we use HHO; a meta-heuristic algorithm that has been inspired by the cooperative hunting mechanism of the Harris Hawk. The experiment will be conducted using HHO to determine the optimal values of the hyperparameters of XG-Boost to improve accuracy and convergence rates compared to grid search and random search. Our preliminary results demonstrate that the HHO-XG-Boost hybrid model could vastly enhance the accuracy and efficiency of the machine learning tasks.

VIII. ACCURACY

Our study shows that a new process is highly dependable having an accuracy level of 96%. It thus surpasses existing methods that only achieve 85% – 90%.

```
In [20]: model = XGBClassifier(
             n_estimators=500,  # Number of boosting rounds (trees)
             max_depth=5,       # Maximum depth of each tree
             learning_rate=0.1, # Step size shrinkage to prevent overfitting
             random_state=42    # Random seed for reproducibility
         )

         # Train the model
         model.fit(X_train_selected, y_train)

         # Predict on the test set
         y_pred = model.predict(X_test_selected)

         # Calculate accuracy
         accuracy = accuracy_score(y_test, y_pred)
         print(f"Accuracy: {accuracy:.2f}")

         Accuracy: 0.96
```

Fig. 4: Accuracy Score

CONCLUSION

We can say that now the email is the crucial and most important tool in the modern world. Messages, homeland to homeland, can now be sent with the aid of the internet. Even though email is one of the most widely used communications businesses, daily, more than 270 billion emails are exchanged, and unfortunately, 57% of them are junk emails (spam). Spam emails serve the same purpose but are known as non-self too. They are the type of commercial or malicious emails that lead to exposing not only a person's confidential information such as banking details, money-making, or anything destructive but also jeopardize groups, individuals, and corporations. In addition to advertisements, such

emails may also include links to phishing or malware-hosting sites that intend to capture kind and protect sensitive personal information, such as bank details. Spam does not only annoy end-users but also poses fraudulent and security risks. So, the designed system is developed to point out and stop unwanted and unsolved communications, which ultimately lead to spam email messages in a lesser quantity. This will surely bring about a great positive impact on everyone – individuals as well as business owners. The future might bring a situation where the given algorithm could be replaced by others and where the introduced feature might be numerous.

REFERENCES

[1] Mohammad Tubishat 1, Feras Al-Obeidat ET AL., "Improved Dandelion Optimizer Algorithm for Spam Detection" 2023.

[2] ALA' M. AL-ZOUBI 1, ANTONIO M. MORA ET AL.,2023 "Multilingual Spam Reviews Detection Based on Pre-Trained Word Embedding and Weighted Swarm Support Vector Machines"

[3] Ashraf S. Mashaleh , Noor Farizah Binti Ibrahim et al.,2022 "Detecting Spam Email with Machine Learning Optimized with Harris Hawks optimizer (HHO) Algorithm".

[4] Hidayet Takci, Fatema Nusrat,2022 "Highly Accurate Spam Detection with the Help of Feature Selection and Data Transformation". https://www.iajit.org/paper/1938/Highly-Accurate-Spam-Detection-with-the-Help-of-Feature-Selection-and-Data-Transformation

[5] Na Song & Yunpeng Ma 2022 "Harris hawks optimization based on global cross-variation and tent mapping". https://link.springer.com/article/10.1007/s11227-022-04869-7

[6] Simran Gibson , Biju Issac ET AL.,2020 "Detecting Spam Email With Machine Learning Optimized With Bio-Inspired MetaheuristicAlgorithms".https://www.researchgate.net/publication/347267306_Detecting_Spam_Email_With_Machine_Learning_Optimized_With_BioInspired_Metaheuristic_Algorithms

[7] D. Dua and C. Graff, "{UCI} Machine Learning Repository." 2017, [Online]. Available:

[8] S. Mirjalili, "Dragonfly algorithm: a new meta-heuristic optimization technique for solving single-objective, discrete, and multi-objective problems," *Neural Comput. Appl.*, vol. 27, no. 4, pp. 1053–1073, 2016, doi: 10.1007/s00521-015-1920-1.

[9] R. V. Rao, V. J. Savsani, and D. P. Vakharia, "Teaching-learning-based optimization: A novel method for constrained mechanical design optimization problems," *CAD Comput. Aided Des.*, vol. 43, no. 3, pp. 303–315, 2011, doi: 10.1016/j.cad.2010.12.015.

[10] Faramarzi, M. Heidarinejad, B. Stephens, and S. Mirjalili, "Equilibrium optimizer: A novel optimization algorithm," *Knowledge-Based Syst.*, vol. 191, p. 105190, 2020, doi: 10.1016/j.knosys.2019.105190.

[11] G. Dhiman and V. Kumar, "Seagull optimization algorithm: Theory and its applications for large-scale industrial engineering problems," *Knowledge-Based Syst.*, vol. 165, 2018, doi: 10.1016/j.knosys.2018.11.024.

[12] Faramarzi, M. Heidarinejad, S. Mirjalili, and A. H. Gandomi, "Marine Predators Algorithm: A nature-inspired metaheuristic," *Expert Syst. Appl.*, vol. 152, no. 113377, pp. 1–48, 2020, doi: 10.1016/j.eswa.2020.113377.