Journal of Intelligent Systems and Internet of Things
Vol. 18, No. 01, PP. 34-47, 2026
DOI: https://doi.org/10.54216/JISIoT.180103

Enhancing Intrusion Detection System Transparency Using SHAP-Driven Support Vector Machine Tuned by Harris Hawks Optimization

Noor Flayyih Hasan[1,*]
[1]Department of Accounting Techniques, Thi-Qar technical College, Southern Technical University, Iraq
Email: noor.f.hasan@stu.edu.iq

Abstract

Due to the increasing prevalence of network attacks, maintaining network security has become significantly more challenging. An Intrusion Detection System (IDS) is a critical tool for addressing security vulnerabilities. IDSs play a vital role in monitoring network traffic and identifying malicious activities. However, two major challenges hinder IDS performance: data imbalance, which weakens the detection of minority class attacks, and overfitting in traditional classifiers such as Support Vector Machines (SVM). This study proposes a novel and transparent IDS framework that integrates several advanced techniques: Variational Autoencoder (VAE) for data augmentation, Mutual Information-based feature selection, Harris Hawks Optimization (HHO) for hyperparameter tuning of the SVM, and SHAP (SHapley Additive exPlanations) for interpretability. VAE is utilized to generate synthetic instances for minority classes, effectively addressing class imbalance. Feature selection is employed to reduce dimensionality and enhance generalization performance. The HHO algorithm is used to adaptively tune the hyperparameters of the SVM, thereby optimizing classification accuracy while mitigating overfitting. Finally, SHAP values are employed to interpret the SVM's decisions, enhancing the transparency and trustworthiness of the system. Experimental evaluations conducted on two benchmark IDS datasets, UNSW-NB15 and NSL-KDD, demonstrate that the proposed VAE-HHO-SVM framework outperforms existing models in terms of accuracy, robustness, and interpretability. The results confirm the effectiveness of combining optimization, explainable AI, and data balancing strategies in modern IDS development. Specifically, the proposed method achieves an accuracy of 98.42% on the NSL-KDD dataset and 97.45% on the UNSW-NB15 dataset—an improvement of 3.17% over other methods.

Received: March 01, 2025 Revised: June 01, 2025 Accepted: July 09, 2025

Keywords: Harris Hawks Optimization; Hyper-parameter optimization; Intrusion Detection Systems; Support vector machine; Variational Autoencoder

1. Introduction

Because of the growing attacks on services based on machines and networks, cybersecurity has become an important subject for protecting systems from threats at a local and global scale over the past decades. However, data encryption and network firewalls presented the main security for computers and networks and satisfied basic security needs, yet broad threats exist that have gone unnoticed and have increased detrimental impacts on the services in total [1]. IDS became the cornerstone in the defense in contrary to cyber threats, having an important role in controlling network traffic and recognizing suspicious/bad functions which can compromise computer systems' accessibility, integrity, and confidentiality [2]. By developing network data volume and complexity, traditional IDS strategies based on rules struggle to keep pace with new and important methods of attack [3]. Accordingly, ML methods were broadly adopted for increasing IDS diagnosis abilities through learning models from historical data [4].

In spite of their success, IDS based on ML meets considerable issues. A basic concern refers to the class imbalance issue: network datasets normally include a huge normal traffic ratio in comparison with relatively few different attack kinds [5]. Such an imbalance causes classifiers to be biased towards the majority class, causing weak diagnosis ratios for rare but crucial intrusions. Also, a lot of models of ML, such as Support Vector Machines (SVM), are prone to overfitting, particularly when hyperparameters are not optimally tuned/while extra and unrelated features exist [6]. Overfitting decreases IDS's generalization ability, leading to degraded performance on unobserved data.

For dealing with such issues, data augmentation techniques were developed for balancing datasets synthetically, and methods of feature selection were used for developing classifier strength [7]. Although a lot of present strategies either lack a principled generative model for augmentation/do not adequately consider hyperparameter tuning and interpretability. The latter is becoming increasingly essential as security analysts need transparent and explainable models to trust automatic IDS decisions. The challenge of class imbalance leads IDS to be biased to the majority class, causing a weak rate of diagnosis for rare, however crucial kinds of attack. In addition, conventional SVM classifiers deal with optimum hyperparameter selection, causing overfitting and decreased diagnosis performance. Such issues hinder the practical, reliable, and interpretable IDS frameworks in real-life scenarios. Considering such challenges, a pressing requirement exists for an IDS framework that not only efficiently balances the dataset but also develops classifier generalization via optimum parameter tuning and presents transparency in its decision-making process for gaining cybersecurity analysts' trust.

Here, its present a general and transparent framework that considers such issues in a unique way. Firstly, the use a Variational Autoencoder (VAE) [8] for creating synthetic minority attack class examples, efficiently mitigating class imbalance with a strong probabilistic generative model. Then, a Mutual Information-based feature selection [9] stage decreases input dimensionality and chooses the most related features, developing classifier accuracy and decreasing overfitting. Accordingly, its develop Harris Hawks Optimization (HHO) [10] for adaptively tuning SVM classifier hyperparameters, optimizing diagnosis performance. At last, it combine SHapley Additive exPlanations (SHAP) [11] for interpreting and explaining SVM's predictions, developing system transparency and trustworthiness. The presented framework is assessed on 2 popular IDS benchmark datasets: UNSW-NB15 [12] and NSL-KDD [13], showing greater robustness, interpretability, and accuracy in comparison with present techniques. The present article makes the main contributions to the IDS domain:

1. To offer VAE usage for creating high-quality synthetic examples for minority attack levels, efficiently considering the class imbalance issue prevalent in IDS datasets.
2. To combine the feature selection step, leveraging shared info for recognizing and retaining the most related features, so decreasing dimensionality, mitigating overfitting, and developing classifier performance.
3. For use HHO metaheuristic mechanism for adaptively optimizing SVM classifier's hyperparameters, developing diagnosis of accuracy and robustness through conventional tuning techniques.
4. To incorporate SHAP values for interpreting and explaining the SVM model decision-making process, developing transparency, and helping cybersecurity professionals in comprehending and trusting the alerts of the system.

This paper is outlined as follows: Part 2 reviews the relevant literature; Part 3 details the offered method; Part 4 shows experimental outcomes and discusses interpretability and implications; Part 5 concludes the study and offers future research directions.

2. Related works

In the last few years, considerable study volume was directed to increasing IDS abilities via machine learning (ML), deep learning (DL), and bio-inspired optimization methods' combination. Networked systems' IOT emergence as well as proliferation accentuated smart IDS frameworks requirement which are scalable, appropriate, and effective. Present part shows the considerable recent research in this field, highlighting the methods developed, their presented strategies, and their restrictions' benefits.

Raghunath et al. [14] employed an IDS tailored for the Internet of Things (IoT), applying ML and feature selection approaches. Their system, which leveraged the NSL-KDD dataset, developed Principal Component Analysis (PCA) for dimensionality reduction, and also applied classifiers such as Random Forest, SVM, and Linear Regression. Their work's basic benefit depends on PSO usage for optimization, resulting in the developed precision and recall. Although dependence of research on the NSL-KDD dataset, which has known restrictions in showing new traffic models, might hinder outcomes' generalizability to real-life areas of IoT.

Alotaibi et al. [15] presented a multiple bio-inspired metaheuristic model integrating Grey Wolf Optimization (GWO) and Quantum Binary Bat Algorithm (QBBA) for feature selection in IDS. They used ML classifiers such as RF, Naive Bayes, and K-Nearest Neighbors (KNN) for assessing the chosen features. Such multiple

optimization decreased features' number to 12, developing computational effectiveness and diagnosis performance. In spite of such strengths, the model's complexity and reliance on hybrid optimization layers might pose challenges of scalability for large-scale developments.

Waghmode et al. [16] considered traditional IDS scalability and false alarm issues through offering a supervised ML framework applying Quantum-inspired Least Squares Support Vector Machine (LS-SVM). Their model obtained min training and testing times, making it appropriate for real-life apps. Therefore, the exhaustive process of feature selection is computationally costly and might not be possible for huge datasets/online systems.

Kumar [17] presented a new IDS given the Linear Discriminant Analysis (LDA) for dimensionality decrease, pursued by the Enhanced Whale Optimization Algorithm (EWOA) for optimum feature selection. They used an ensemble classifier integrating RF and XGBoost. Such an ensemble strategy takes advantage of developed classification robustness. Although, framework's performance was just confirmed on 1 dataset, raising questions on its generalizability over various network scenarios.

Kanna and Santhi [18] defined the scalable and flexible multiple IDS model applying a MapReduce-driven framework with a Black Widow Optimization (BWO)-tuned Convolutional Long Short-Term Memory (CONV-LSTM) network. Feature selection was carried out through the mechanism of Artificial Bee Colony (ABC). The presented model obtained high accuracy on hybrid datasets, with decreased time of calculation and false positives. The basic restriction refers to computational overhead defined by multiple DL models that may not be practical for a real-life IDS with no high-performance computing resources.

Qiu et al. [19] provided a new CNN–Decision Tree (CNN-DT) multiple model for IDS, developed via a uniform multiple pooling approach and optimized by the Actor-Critic deep reinforcement learning mechanism. Their technique is performing better than the non-optimized version. Decision trees usage presents interpretable decisions when CNNs guarantee strong feature extraction. Although the model's dependency on reinforcement learning for hyperparameter tuning increases training complexity and needs of source.

Alshinwan et al. [20] presented multiple techniques of optimization known as PDO-DE that integrates Prairie Dog Optimization and Differential Evolution for feature selection and parameter tuning. However, particular metrics of performance were not completely detailed in the presented conclusion, this combination targets at developing IDS models' convergence rate and accuracy. The basic advantage is a more explorative search space ability; however, the multiple aspects of the mechanism might define issues in the case of parameter calibration and convergence stability.

Abualigah et al. [21] presented a new feature selection technique for IDS in Wireless Sensor Networks by combining SVM with a modified Aquila Optimizer (mAO). Such a strategy was confirmed on the KDD'99 dataset and assessed via different performance metrics such as false alarm rate, accuracy, execution time, diagnosis rate, and number of selected features. Among its basic benefits, its developed computational efficiency, high accuracy, and low false alarm rate were due to feature dimensionality decrease. Although the considerable restriction is its dependence on the outdated KDD'99 dataset, which might hinder its applicability to new and more complicated areas of the network. In addition, the model's performance under real-life/highly imbalanced situations was not evaluated.

Babu and Rao [22] defined the developed IDS given the DL, using the Attention-based Nested U-Net (ANU-Net) framework for dealing with concerns like data imbalance and rare attack diagnosis. Their methods included 3 preprocessing stages—duplicate removal, label transformation, and data normalization—pursued by feature selection applying the Improved Flower Pollination Algorithm (IFPA) and hyperparameter optimization through the Improved Monarchy Butterfly Optimization (IMBO) mechanism. Also, parallel computing usage via the MapReduce framework has developed computational pace. This technique's main advantage is its efficient control of imbalanced datasets, strong optimization processes, as well as scalability to huge datasets. So, its complexity and dependence on hybrid metaheuristics might have issues of implementation and require noticeable computational resources.

Ponmalar and Dhanakoti [23] mentioned conventional IDS restrictions in huge areas of data through offering multiple model that integrates ensemble SVM with the Chaos Game Optimization (CGO) mechanism. Such a mechanism was modelled for controlling heterogeneous and voluminous security data. The benefits of the model depend on its suitability for big-scale data analysis, high classification accuracy, and reduced false positives. However, its evaluation on a single dataset might restrict its generalizability. The strategy of ensemble SVM, when efficient, can be computationally demanding and less interpretable in comparison with simpler models.

Nandhini and SVN [24] used the Principal Component Analysis (PCA) technique to minimize the dimensionality of the dataset. The Improved Harris Hawks Optimizer (IHHO) is used for effective feature selection, resulting in significant global search capabilities. A two-stage classifier is proposed for classification, with the first stage using Support Vector Machine (SVM) and the second stage using K-Nearest Neighbors.

Aswanandini and Deepa [25] created a big data analytics model to classify intrusion datasets using Optimized SVM. In this model, a hybrid technique called Hyper-Heuristic Particle Swarm Optimization (HHPSO) is used to optimize the SVM configuration. This can dramatically reduce model complexity while also reducing training time. To achieve this goal, hyper-heuristic optimization is paired with Particle Swarm Optimization (PSO) to optimize the margin parameter, kernel type, and kernel parameter to improve accuracy and reduce model complexity in the SVM model.

The reviewed researches collectively show the ability of combining ML, DL, and bio-inspired optimization methods for increasing IDS performance and reliability. When a lot of such techniques illustrate high diagnosis accuracy and decreased false alarm rates, restrictions like sensitivity to hyperparameter tuning persist, computational complexity, and lack of generalizability exist. Such outcomes bold the demand for ongoing study into light, interpretable, and scalable IDS frameworks which keep strong performance over varied and evolving network areas.

3. Proposed method

For efficiently mentioning main issues in IDS—known as lack of model interpretability, data imbalance, overfitting, and inefficient parameter tuning—to offer a new and general architecture. Such a technique combines 4 basic elements: data augmentation through Variational Autoencoder (VAE), feature selection applying Mutual Information (MI), hyperparameter optimization via Harris Hawks Optimization (HHO), and model explainability applying SHAP values. Every module contributes to uniformly developing SVM classifier transparency, robustness, and accuracy, which serves as the main predictive engine in the system. The VAE-HHO-SVM model for IDS is illustrated in Figure 1.

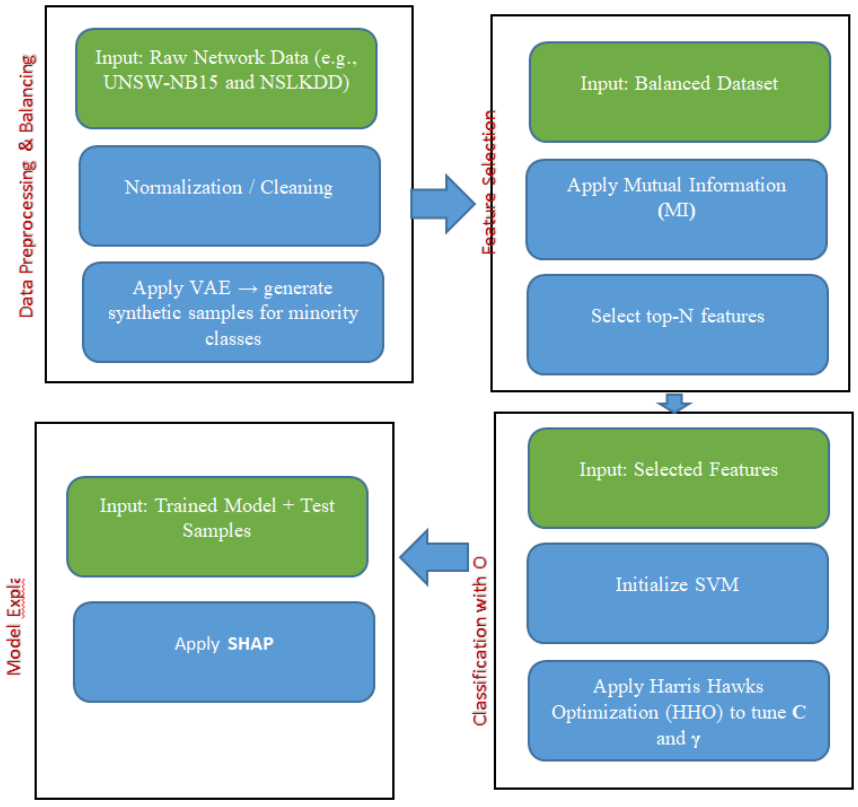


Figure 1. VAE-HHO-SVM in intrusion detection

3.1 Normalization

The technique of normalization decreases the variety in input features, given the min and max values. The classifier model could efficiently learn the features because of the less difference in the input data. The min-max normalization formula is provided in Equation (1).

(1)

$$x = \frac{x - x_{min}}{x_{max} - x_{min}}$$

The xN is normalized data, x is the input value, $xmin$ is min value in the feature, and $xmax$ is the maximum value in the feature.

3.2. Variational Autoencoder (VAE) for Data Augmentation

One of the fundamental challenges in intrusion detection is the severe class imbalance in network traffic datasets, where attack samples are significantly outnumbered by normal traffic instances. To mitigate this, it employ a Variational Autoencoder (VAE) to generate synthetic samples for the minority attack classes.

The VAE is a probabilistic generative model that learns the underlying distribution of input data by encoding it into a latent space and then decoding it back to reconstruct the data. By sampling from the learned latent space, VAE can produce diverse and realistic synthetic samples that augment the training dataset. This approach not only balances the dataset but also preserves the complex feature distributions of minority classes, improving the classifier's ability to detect rare intrusions.

The autoencoder is a neural net that maps its input to a result of completely same dimension. In the middle of the neural net bottleneck exists, which is the uniform autoencoder feature. They could aid in decreasing noise/getting a lower-dimensional input representation. The particular autoencoder variation is a variational autoencoder. Whole autoencoders have an encoder, followed by a latent space, and a decoder. The encoder compresses the input x into a latent space z, and a decoder decodes from this latent space z to get the reconstruction x̂. The encoder and decoder can be interpreted as an identification and generative model, in turn.

(2)
$$z \approx \mathrm{Enc}(x) = q_{\varnothing}(z|x),\ \hat{x} \approx \mathrm{Dec}(z) = p_{\theta}(x|z),$$

That qØ(z|x) is the true posterior of pθ(x|z), and θ is applied for showing encoder and decoder models, in turn. For an autoencoder, the latent space is different, so sampling from this space would not cause anything meaningful. A variational autoencoder takes such various values and attempts to recognize a shared, familiar latent space. So, the variational autoencoder not only attempts to rebuild its input, but also attempts to shape the space in the latent space. The loss function L (Eq. (2)) includes 2 terms; the first term penalizes the error of rebuilding among input x and the output x̂, and the second term penalizes the error between the previous pθ(z) and the learned share qØ(z|x).

(3)
$$L = \log p_{\theta}(x|z) - D_{KL}(q_{\varnothing}(z|x)||p_{\theta}(z))$$

In Equation (2), pθ(z) is the previous share, and DKL is the Kullback-Leibler divergence that could scale how much info is lost while q is applied for showing p.

3.2.1. Encoder model of VAE

It took 2 two-layer densely connected encoder model for the VAE. The next 2 layers after the input x are the decoder. The number selected was (24, 24). The experimented with lower numbers like (10,10), (15,15), (20,20), and (24,20). The best latent space visualization was achieved for (24,24).

3.2.2. Decoder model of VAE

The decoder takes the latent layer code and attempts to rebuild basic crash data. For the decoder, it had 24 neurons for the first layer and a similar number for the second layer. A sigmoid activation function was applied at the end of the decoder. A layer of output layer would be similar to the input layer, which is 24. Whole neurons were densely connected.

3.3. Feature Selection Using Mutual Information

High-dimensional network data sometimes includes extra and unrelated features that could degrade performance and enhance classifiers' complexity. For developing model generalization ability and decreasing overfitting, to incorporate the stage of feature selection given the Mutual Information (MI).

MI scales dependency among every feature and the target class, making us able to choose features that transfer the most predictive info on kinds of intrusion. It decreases dimensionality, develops computational efficiency, and improves classifier robustness.

MI among feature X and the class label Y scales their dependency:

(4)
$$I(X;Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

That p(x,y) is the joint probability share, p(x), p(y) are marginal shares.

Features are ranked by their MI scores with the class, the top-ranked features are chosen for designing later.

3.4 Harris Hawks Optimization (HHO) for SVM Hyperparameter Tuning

SVMs are robust classifiers widely applied in IDS because of their efficiency in high-dimensional spaces. Although their performance completely relies on hyperparameter selection, like penalty parameter C and kernel parameters.

For optimizing such hyperparameter exploration-exploitation balance and quick convergence. HHO repeatedly looks for the space of hyperparameters for finding optimum amounts that enhance classification accuracy, so decreasing the risk of overfitting and improving the performance of diagnosis.

Heidari et al. [26] improved the mechanism known as HHO (Harris hawks' optimization). This is taken from the hunting type and Harris's hawks' cooperation. Several hawks collaborate while attacking their prey from various directions to surprise and disable it disabled. So, for helping in various hunting approaches' selection, this is related to different scenery and types of prey flying. Exploring a prey, transitioning from exploration to exploitation as well and exploitation are the 3 basic HHP stages. In this diagram, the whole HHO process is shown. Each stage diagram is provided below.

3.4.1. Exploration Phase

It is mathematically designed basically for searching, prey diagnosis, and waiting. Harris's hawks are the alternative/best at each stage. Harris's hawks' position X(i + 1) could be formulated based on Equ. (5):

(5)
$$X(i+1) = \begin{cases} (X_{rand}(i)) - r_1|X_{rand}(i) - 2r_2X(i)| & if\ q \geq 0.5 \\ (X_{rabbit}(i) - X_m(i)) - badrr_3(LB + r_4(UB - LB)) & if\ q < 0.5 \end{cases}$$

That is the present repetition, Xrabbit is the rabbit's position, Xrand is a randomly selected hawk from the present population, rj, j = 1, 2, 3, 4, q are random numbers among 0 and 1, and Xm is the medium hawks' position that could be computed using:

(6)
$$X_m(i) = \frac{1}{N}\sum_{j=1}^{N} X_j(i)$$

That vector Xj shows every hawk j's position, N is hawks' number.

3.4.2. Transition from Exploration to Exploitation

The HHO alternates between exploration and exploitation based on the rabbit's escaping energy. In addition, the rabbit's energy could be computed by applying the following formula:

(7)
$$E = 2E_0\left(1 - \frac{i}{T}\right)$$

That E shows the rabbit's escaping energy, T shows max iterations' size, and E0 ∈ (−1, 1) shows basic energy at every stage.

(8)
$$E_0 = 2\,rand(\,) - 1$$

The HHO can assign rabbit condition given the direction of E0 (the HHO enters the exploration step for locating the prey when |E| ≥ 1, otherwise, in stages of exploitation, this approach looks for exploiting solutions' proximity).

3.4.3. Exploitation Phase

Here, hawks besiege the prey from whole directions to hunt it, the siege is hard/soft based on the prey's energy. In this siege, the prey's escape relies on selection r (which succeeds in escaping when r < 0.5). In addition, when |E| ≥ 0.5, the HHO is besieging softly; otherwise, this is besieging hard. Based on prey escape and hawks–hawks' approaches in pursuit events, HHO performs 4 attack approaches: a hard siege, a soft siege, a hard siege with progressive quick dives, and a soft siege with progressive quick dives. Especially, the rabbit has sufficient energy for escaping when |E| ≥ 0.5; although, the prey's capability for escaping /not relies on the two |E| and r amounts.

Soft Siege (|E| ≥ 0.5 and r ≥ 0.5)

The present way could be considered as:

(9)

(10)
$$X(i+1) = \Delta X(i) - E|JX_{abbit}(i) - X(i)|$$
$$\Delta X(i) = X_{rabbit}(i) - X(i)$$

That ΔX(i) shows the difference between the rabbit's current place and the rabbit's place vector at the i iteration, J = 2(1 − r5) is the rabbit's random jumping intensity in the escape process, r5 ∈ (0, 1) is a random number.

Hard Siege (|E| < 0.5 and r ≥ 0.5)

Here, present locations could be updated with the formula below:

$$X(i+1) = X_{rabbit}(i) - E|\Delta X(i)|$$

Soft Siege with Progressive Rapid Dives (r < 0.5 and| E| ≥ 0.5)

As for the soft siege, hawks decide their next move with the equation below:

$$Y = X_{rabbit}(i) - E|JX_{rabbit}(i) - X(i)|$$

The hawks dive based on the laws below, according to LF-driven models:

$$Z = Y + S \times LF(D)$$

Where D shows the issue dimension, S1×D shows a random vector. The Levy flight (LF) could be computed by Eq. (14):

(14)
$$LF(D) = 0.01 \times \frac{\mu \times \sigma}{|v|^{\frac{1}{\beta}}}, \sigma = \left(\frac{\Gamma(1+\beta)\sin\left(\frac{\pi\beta}{2}\right)}{\Gamma\left(\frac{1+\beta}{2}\right) \times \beta \times 2^{\left(\frac{\beta-1}{2}\right)}} \right), \beta = 1.5$$

That μ and v show a range of random numbers between 0 and 1. So, Equ. (15) could be applied to describe the last approach of this step, which is to update the hawks' conditions:

(15)
$$X(i+1) = \begin{cases} Y & if\ F(Y) < F(X(i)) \\ Z & if\ F(Z) < F(X(i)) \end{cases}$$

Hard Siege with Progressive Rapid Dives (r < 0.5 and |E| < 0.5). The hawk is always near its prey in this stage. Formulas below could be applied for computing Y and Z:

(16)
$$Y = X_{rabbit}(i) - E|JX_{rabbit}(i) - X_m(i)|$$

(17)
$$Z = Y + S \times LF(D)$$

where

(18)
$$X_m(i) = \frac{1}{N}\sum_{i=1}^{N} X_i(i)$$

This work's basic aim was to calculate the SVM parameter by applying the HHO mechanism for effectively grouping traffic data.

3.5. Explainable AI via SHapley Additive exPlanations (SHAP)

Understanding and trusting IDS decisions are critical for cybersecurity analysts. For presenting transparency, to develop SHapley Additive exPlanations (SHAP), a new explainability technique based the cooperative game theory.

SHAP calculates every attribute contribution to the last SVM classifier prediction, making decision rationale visualization and interpretation possible. It facilitates the recognition of crucial attributes facilitated causing particular IDS and developing entire system trustworthiness.

SHAP values are obtained from cooperative game theory, attributing every attribute contribution to the prediction. The Shapley value ϕi for feature i is calculated as:

(19)

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right]$$

That N refers to all features' collection, S is a subset excluding feature i, fS is the model trained on subset S, xS are the feature values in S. SHAP decomposes the output model into additive feature contributions, making a detailed perspective able into which features drive particular predictions.

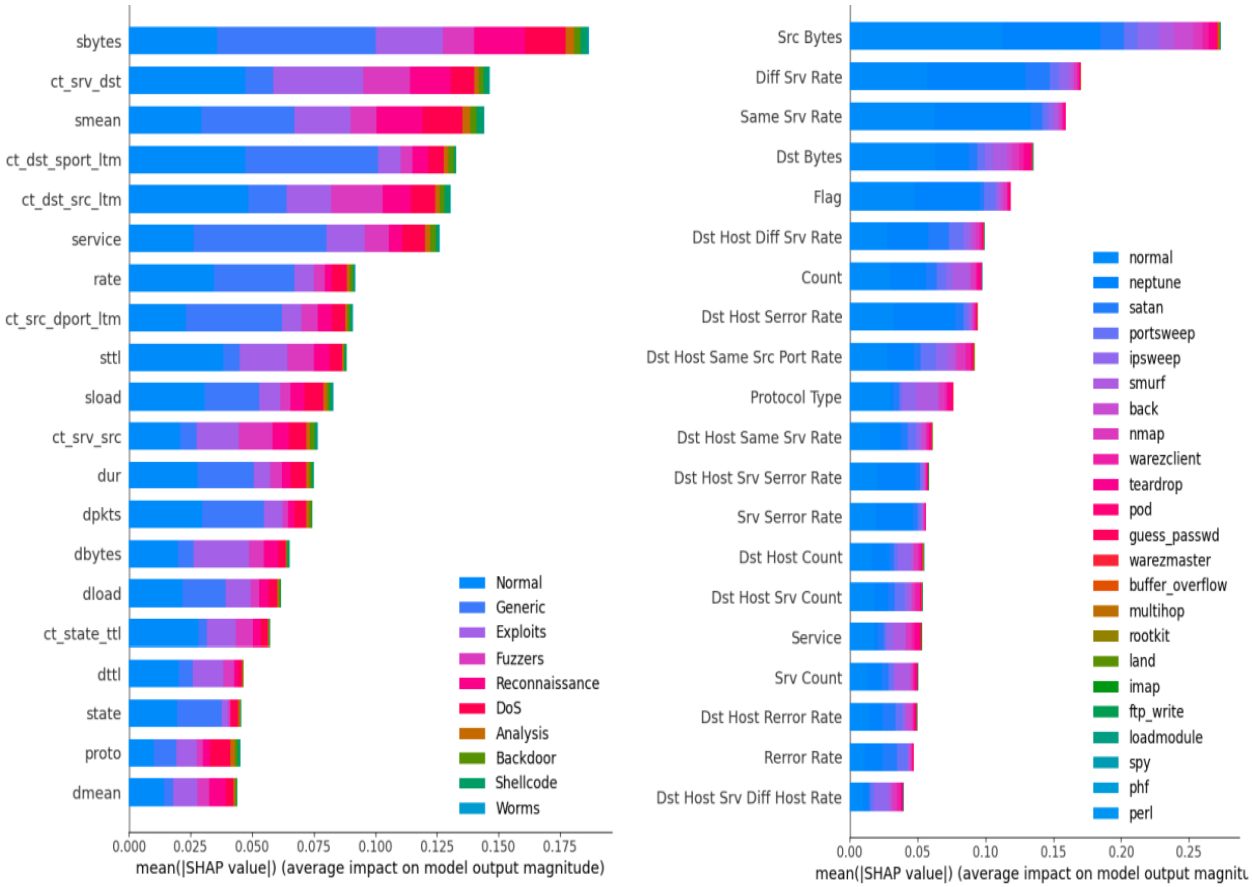


Figure 2. (a) Global SHAP values for the UNSW-NB15 dataset. (b) Global SHAP values for the NSL-KDD dataset

4. Results and Discussions

4.1 Experimental Setting

For assessing the presented IDS strategy efficiency, whole tests were performed applying Python 3.x on a machine with an Intel Core i7 processor, 16 GB RAM, and an NVIDIA GPU (in case it is accessible) for accelerating calculations of DL. The models were performed applying ML and DL libraries like Scikit-learn, TensorFlow, and Keras. The training step used stratified 5-fold cross-validation to guarantee outcomes' robustness and generalizability.

4.2 Datasets

In this study used two benchmark datasets that were used previously in several studies. Datasets: The dataset description is provided as:

UNSW-NB15 dataset [12]: Labeling features of 47 features are characterized in every record. The real-life traffic network packets have 47 features and network access of labelling features as usual/unusual. The 5 sets are grouped in 47 features: flow, additional generated features, basic, time, and content features.

NSL-KDD [13]: The NSL-KDD dataset is obtained from the KDD 99 dataset by removing extra and duplicate records, which is more reasonable in the structure and size of the data. 67343 number of normal data samples, 125973 number of sample data, 42 features, 118191 number of attacks.

Table 1: Features Selected by MI

NSL-KDD	UNSWNB15
duration	dur
protocol_type	service
service	state
Flag	spkts
src_bytes	dpkts
dst_bytes	sttl
wrong_fragment	dttl
num_failed_logins	sload
logged_in	dload
num_compromised	sloss
root_shell	dloss
su_attempted	sinpkt
num_root	dinpkt
num_shells	swin
num_access_files	stcpb
num_outbound_cmds	dtcpb
is_host_login	dwin
srv_count	tcprtt
serror_rate	synack
rerror_rate	ackdat
same_srv_rate	smean
diff_srv_rate	dmean
srv_diff_host_rate	trans_depth
dst_host_srv_count	response_body_len
dst_host_same_srv_rate	ct_srv_src
dst_host_diff_srv_rate	ct_state_ttl
dst_host_srv_diff_host_rate	ct_dst_ltm
dst_host_srv_serror_rate	ct_src_dport_ltm
dst_host_rerror_rate	ct_dst_sport_ltm
dst_host_srv_rerror_rate	ct_dst_src_ltm
	is_ftp_login
	ct_ftp_cmd
	ct_flw_http_mthd
	ct_src_ltm
	ct_srv_dst
	is_sm_ips_ports

4.3 Evaluation Metrics

Precision, Recall, accuracy, and F-score were applied for assessing IDS (F). A low false alarm rate, high precision, and a high diagnosis rate are needed for the result. Such features are calculated by applying a confusion matrix. True Positive (TP) is the number of assault reports that were accurately grouped in the confusion matrix. The number of accurately recognized usual data is called True Negative (TN). The medium of normal records, which were wrongly grouped, is called false positives (FP). False Negative (FN) is the number of inappropriate assault records.

Accuracy: shows the real diagnosis rate in the whole traffic trace. It is achievable as:

$$\text{Accuracy} = \frac{\text{TN} + \text{TP}}{\text{TN} + \text{FP} + \text{TP} + \text{FN}} \tag{20}$$

Precision: shows certain intrusions in comparison with those predicted by a NIDS. It might be decreased that the higher the PR, the smaller the false alarm:

$$Precision = \frac{TP}{TP + FP} \tag{21}$$

Sensitivity/Recall: such a metric scales the links' rate to successfully grouped anomalies. This illustrates how well the model is at picking up abnormalities from each anomalous connection. This is achievable as:

$$\text{Recall} = \frac{\text{TP}}{\text{FN} + \text{TP}} \tag{22}$$

F-measure (F1-score): It is a performance metric applied for assessing classification model accuracy, particularly where the data is not balanced. This is the Precision and Recall harmonic concept; this presents a balance between the two.

$$Fmeasure = 2 \times \frac{Precision \times Sensitivity}{Precision + Sensitivity} \tag{23}$$

Also, the Receiver Operating Characteristic (ROC) Curve and Area Under the Curve (AUC) were applied for visualizing and quantifying classifier performance.

4.3.1 Performance on UNSW-NB15 Dataset

On the UNSW-NB15 dataset, the presented technique showed greater performance over the whole metrics of assessment. The model obtained a precision of 95.84%, an F1-score of 96.17%, accuracy of 97.45%, recall of 96.51%. The false positive rate was considerably low, bolstering the model's reliability in real-life areas of security. The presented multiple strategies were especially efficient in recognizing minority class attacks like Shellcode and Backdoor that are normally underrepresented in the dataset. In comparison with baseline models such as traditional SVM and RF, the technique illustrated a performance improvement of over 3.17% in accuracy.

Table 2: Performance of the proposed method on UNSW-NB15 Dataset

Method	Accuracy	Recall	Precision	F1-score
RF	94.28%	93.70%	92.85%	93.27%
SVM	92.63%	91.45%	90.12%	90.78%
Proposed method	97.45%	96.51%	95.84%	96.17%

4.3.2 Performance on NSL-KDD Dataset

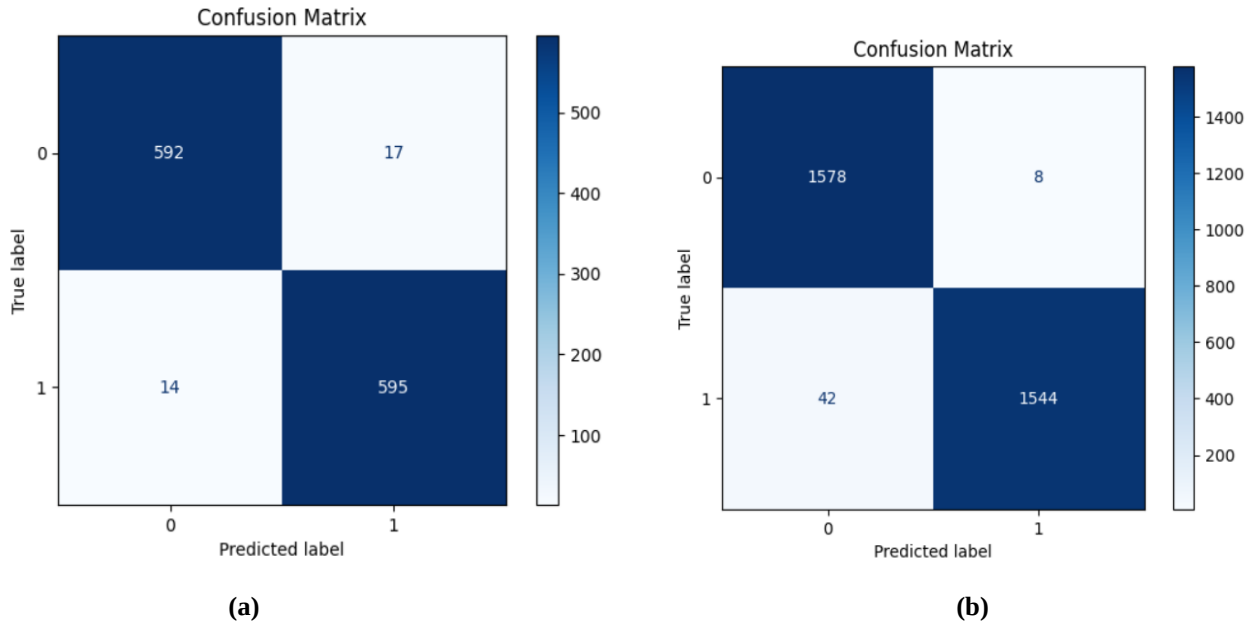
While assessed on the NSL-KDD dataset, the presented model kept its high performance with an entire precision of 97.89%, a recall of 98.23%, an F1-score of 98.06%, accuracy of 98.11%. It successfully diagnosed the wide attack groups, showing rare kinds such as U2R and R2L, which a lot of models attempt to misclassify. In comparison with present techniques like strategies based on CNN, RNN, the model showed better generalization and strength, especially in decreasing false alarms and diagnosing low-frequency attacks.

Table 3: Performance of the proposed method on the NSL-KDD Dataset

Method	Accuracy	Recall	Precision	F1-score
RNN	95.67%	95.10%	94.85%	94.97%
CNN	93.84%	93.21%	92.73%	92.97%
Proposed method	98.42%	98.23%	97.89%	98.06%

Figure 3 illustrates the confusion matrices for the proposed model evaluated on two benchmark datasets: UNSW-NB15 and NSL-KDD. In subfigure (a), which corresponds to the UNSW-NB15 dataset, the model correctly classified 592 instances as class 0 (True Negatives) and 595 instances as class 1 (True Positives), with only 17 False Positives and 14 False Negatives. This reflects a highly balanced classification performance, with minimal misclassification. Similarly, in subfigure (b), the confusion matrix for the NSL-KDD dataset shows 1578 True Negatives and 1544 True Positives, with only 8 False Positives and 42 False Negatives. These results confirm the robustness and generalizability of the proposed model across diverse intrusion detection datasets. The low number of false classifications in both datasets indicates that the model achieves high accuracy and precision while maintaining a strong balance between sensitivity (recall) and specificity, which are crucial metrics in cybersecurity applications such as phishing or intrusion detection.

Figure 4 presents the Receiver Operating Characteristic (ROC) curves along with the Area Under the Curve (AUC) values for the proposed method evaluated on the UNSW-NB15 and NSL-KDD datasets. In subfigure (a), the ROC curve for the UNSW-NB15 dataset demonstrates an AUC of 0.99, indicating near-perfect classification performance. This high AUC value reflects the model's strong ability to distinguish between phishing and legitimate websites with minimal false positives or false negatives. In subfigure (b), the ROC curve for the NSL-KDD dataset achieves an AUC of 1.00, which denotes an ideal classifier with perfect discrimination between classes. These results highlight the reliability and effectiveness of the proposed SI-WOA framework in identifying malicious behavior across different datasets. The ROC-AUC analysis reinforces the previously reported performance metrics, confirming that the model maintains excellent sensitivity and specificity, which are vital for real-world applications in network security and threat detection.



(a) **(b)**

Figure 3. a) Confusion matrix with UNSW-NB15 dataset. b) Confusion matrix with NSL-KDD dataset

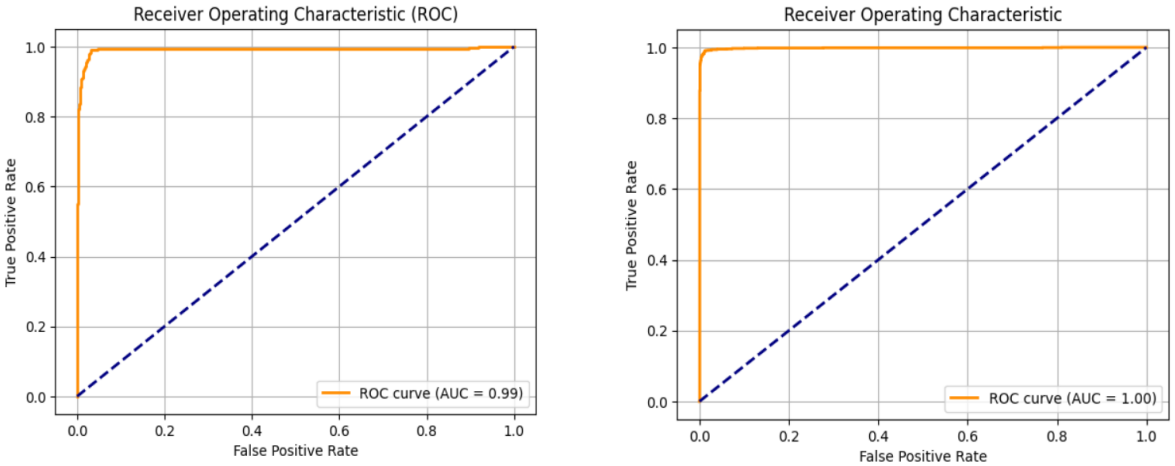


Figure 4. a) ROC with UNSW-NB15 dataset. b) ROC with NSL-KDD dataset

4.4 Comparative Analysis

The general comparison was performed between the presented technique and the recent IDSs. On both datasets, the model consistently outperformed other techniques. Table 4 compares the proposed method to many current methodologies, using the NSL-KDD and UNSW-NB15 datasets. As observed, the proposed method outperforms previous models in terms of accuracy. The suggested approach yields an outstanding 98.42% accuracy on the NSL-KDD dataset, which is significantly higher than the results reported by Nandhini and SVN [24] (95.01%) and Aswanandini and Deepa [25] (91.30%). Similarly, using the UNSW-NB15 dataset, the suggested technique reaches 98.42% accuracy, surpassing the results of Ponmalar and Dhanakoti [23] (96.29%) and Waghmode et al [16]. (93.3%). These improvements demonstrate the robustness and generalization capability of the proposed framework across different intrusion detection datasets. The consistent high performance across both datasets confirms the model's superiority in identifying intrusion activities compared to machine learning approaches. This enhanced performance can be attributed to the integration of VAE-HHO-SVM for adaptive training, which enables more precise learning and better generalization.

Table 4: Comparison of the proposed method with existing methods

Method	Dataset	Accuracy
Nandhini and SVN [24]	NSL-KDD	95.01
Aswanandini and Deepa [25]	NSL-KDD	91.30
Ponmalar and Dhanakoti [23]	UNSW-NB15	96.29%
Waghmode et al. [16]	UNSW-NB15	93.3
Proposed method	NSL-KDD	98.42%
Proposed method	UNSW-NB15	98.42%

5. Conclusion

In this paper, apresented a balanced, new, transparent IDS framework that combines some developed methods for considering the main issues in network security, like hyperparameter tuning, data imbalance, lack of interpretability, and overfitting. The presented system leverages a Variational Autoencoder (VAE) for creating

synthetic instances for minority attack classes, efficiently balancing the training dataset and developing classifier generalization. For decreasing dimensionality and bolding the most informative features, have developed Mutual Information (MI) for feature selection, developing model effectiveness, and reducing noise. For the task of classification, an SVM was adopted because of its strength and ability to control high-dimensional data. For later performance development, the mechanism of Harris Hawks Optimization (HHO) was developed for automatically tuning SVM's hyperparameters. At last, SHapley Additive exPlanations (SHAP) were combined into the pipeline for presenting the two global and local interpretability, making transparency and trust possible in the model's predictions. Experimental assessments on hybrid benchmark IDS datasets like NSL-KDD and UNSW-NB15 showed that in the presented technique performs better than some present strategies in terms of interpretability, accuracy, and reliability. Explainable AI integration into the IDS process is especially worthy for crucial apps where comprehending the rationale behind predictions is as essential as the predictions themselves. For later study, have plan to develop a synthetic data generation process applying conditional VAEs or diffusion-based models, explore deep feature selection applying neural attention algorithms, and investigate ensemble learning approaches for later developing diagnosis performance. Also, developing such a framework in real-life areas and assessing its robustness in contrast with adversarial attacks remains a considerable direction for practical applications.

References

[1] O. Ahmed, "Enhancing Intrusion Detection in Wireless Sensor Networks through Machine Learning Techniques and Context Awareness Integration," *International Journal of Mathematics, Statistics, and Computer Science*, vol. 2, pp. 244–258, 2024. [Online]. Available: https://doi.org/10.59543/ijmscs.v2i.10377.

[2] L. Zolfagharipour, M. H. Kadhim, and T. H. Mandeel, "Enhance the Security of Access to IoT-based Equipment in Fog," in *Al-Sadiq International Conference on Communication and Information Technology (AICCIT) 2023*, Jul. 4, 2023, pp. 142-146. IEEE.

[3] D. Ditale, M. Albanese, K. Sun, and J. Pan, "CyberMALT: Machine Learning-Assisted Traffic Analysis for Cyber Threat Detection and Classification," in *IEEE 22nd Consumer Communications & Networking Conference (CCNC) 2025*, Jan. 10, 2025, pp. 1-6. IEEE.

[4] K. Noor, A. L. Imoize, C. T. Li, and C. Y. Weng, "A review of machine learning and transfer learning strategies for intrusion detection systems in 5 G and beyond," *Mathematics*, vol. 13, no. 7, Art. 1088, Mar. 26, 2025.

[5] L. Zolfagharipour and M. H. Kadhim, "A Technique for Efficiently Controlling Centralized Data Congestion in Vehicular Ad Hoc Networks," *International Journal of Computer Networks and Applications (IJCNA)*, vol. 12, no. 2, pp. 267-277, 2025.

[6] J. Ehmer, Y. Savaria, B. Granado, J. P. David, and J. Denoulet, "Network Attack Classification with a Shallow Neural Network for Internet and Internet of Things (IoT) Traffic," *Electronics*, vol. 13, no. 16, Art. 3318, Aug. 21, 2024.

[7] Y. Zhang, Q. Wu, G. Liu, J. Tian, S. Liang, and C. Zhang, "Research on Structure and Hyperparameter Optimization of Lightweight ViT Model Based on SVM Optimization in Object Detection," in *IEEE 6th International Conference on Civil Aviation Safety and Information Technology (ICCASIT) 2024*, Oct. 23, 2024, pp. 1644-1648. IEEE.

[8] L. Moles, A. Andres, G. Echegaray, and F. Boto, "Exploring data augmentation and active learning benefits in imbalanced datasets," *Mathematics*, vol. 12, no. 12, Art. 1898, Jun. 19, 2024.

[9] B. M. Kessels, R. H. Fey, and N. van de Wouw, "Mutual information-based feature selection for inverse mapping parameter updating of dynamical systems," *Multibody System Dynamics*, vol. 2024, pp. 1-28, Aug. 19, 2024.

[10] A. A. Heidari, S. Mirjalili, H. Faris, I. Aljarah, M. Mafarja, and H. Chen, "Harris hawks optimization: Algorithm and applications," *Future Generation Computer Systems*, vol. 97, pp. 849-872, Aug. 1, 2019.

[11] L. Antwarg, R. M. Miller, B. Shapira, and L. Rokach, "Explaining anomalies detected by autoencoders using Shapley Additive Explanations," *Expert Systems with Applications*, vol. 186, Art. 115736, Dec. 30, 2021.

[12] N. Moustafa and J. Slay, "UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," in *Military Communications and Information Systems Conference (MilCIS) 2015*, Nov. 10, 2015, pp. 1-6. IEEE.

[13] D. D. Protić, "Review of KDD Cup '99, NSL-KDD, and Kyoto 2006+ datasets," *Vojnotehnički glasnik/Military Technical Courier*, vol. 66, no. 3, pp. 580-596, Apr. 8, 2018.

[14] M. P. Raghunath, S. Deshmukh, P. Chaudhari, S. L. Bangare, K. Kasat, M. Awasthy, B. Omarov, and R. R. Waghulde, "PCA and PSO-based optimized support vector machine for efficient intrusion detection in the Internet of Things," *Measurement: Sensors*, vol. 37, Art. 101806, Feb. 1, 2025.

[15] M. Alotaibi, H. A. Mengash, H. Alqahtani, A. M. Al-Sharafi, A. E. Yahya, S. R. Alotaibi, A. O. Khadidos, and A. Yafoz, "Hybrid GWQBBA model for optimized classification of attacks in Intrusion Detection System," *Alexandria Engineering Journal*, vol. 116, pp. 9-19, Mar. 1, 2025.

[16] P. Waghmode, M. Kanumuri, H. El-Ocla, and T. Boyle, "Intrusion detection system based on machine learning using least square support vector machine," *Scientific Reports*, vol. 15, no. 1, Art. 12066, Apr. 8, 2025.

[17] S. V. Kumar, "An enhanced whale optimizer-based feature selection technique with an effective ensemble classifier for a network intrusion detection system," *Peer-to-Peer Networking and Applications*, vol. 18, no. 2, pp. 1-28, Apr. 2025.

[18] P. R. Kanna and P. Santhi, "An enhanced hybrid intrusion detection using mapreduce-optimized black widow convolutional LSTM neural networks," *Wireless Personal Communications*, vol. 138, no. 4, pp. 2407-2445, Oct. 2024.

[19] L. Qiu, Z. Xu, L. Lin, J. Zheng, and J. Su, "Design and Optimization of Hybrid CNN-DT Model-Based Network Intrusion Detection Algorithm Using Deep Reinforcement Learning," *Mathematics*, vol. 13, no. 9, Art. 1459, Apr. 29, 2025.

[20] M. Alshinwan, O. A. Khashan, M. Khader, O. Tarawneh, A. Shdefat, N. Mostafa, and D. S. AbdElminaam, "Enhanced Prairie Dog Optimization with Differential Evolution for solving engineering design problems and network intrusion detection systems," *Heliyon*, vol. 10, no. 17, Sep. 15, 2024.

[21] L. Abualigah, S. H. Ahmed, M. H. Almomani, R. A. Zitar, A. R. Alsoud, B. Abuhaija, E. S. Hanandeh, H. Jia, D. S. Elminaam, and M. A. Elaziz, "Modified Aquila Optimizer feature selection approach and support vector machine classifier for intrusion detection system," *Multimedia Tools and Applications*, vol. 83, no. 21, pp. 59887-59913, Jun. 2024.

[22] K. S. Babu and Y. N. Rao, "Improved Monarchy Butterfly Optimization Algorithm (IMBO): Intrusion Detection Using MapReduce Framework Based Optimized ANU-Net," *Computers, Materials & Continua*, vol. 75, no. 3, Jun. 1, 2023.

[23] A. Ponmalar and V. Dhanakoti, "An intrusion detection approach using ensemble support vector machine based chaos game optimization algorithm in big data platform," *Applied Soft Computing*, vol. 116, Art. 108295, Feb. 1, 2022.

[24] U. Nandhini and S. K. SVN, "An improved Harris Hawks optimizer-based feature selection technique with an effective two-stage classifier for a network intrusion detection system," *Peer-to-Peer Networking and Applications*, vol. 17, no. 5, pp. 2944-2978, Sep. 2024.

[25] R. Aswanandini and C. Deepa, "Network Intrusion Classification using Configuration Optimized Support Vector Machines," in *International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA) 2021*, Oct. 8, 2021, pp. 1-6. IEEE.

[26] E. Heidari, "A novel energy-aware method for clustering and routing in IoT based on whale optimization algorithm & Harris Hawks optimization," *Computing*, vol. 106, no. 3, pp. 1013-1045, Mar. 2024.