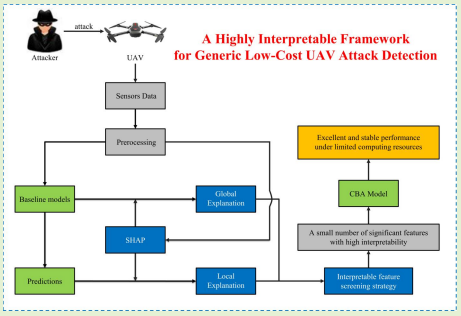
A Highly Interpretable Framework for Generic Low-Cost UAV Attack Detection

Shihao Wu, Yang Li, *Member, IEEE*, Zhaoxuan Wang, Zheng Tan, and Quan Pan, *Member, IEEE*

***Abstract*—The increasing prevalence of cyber-attacks on unmanned aerial vehicles (UAVs) has led to research on effective detection methods. However, current approaches often lack transferability and interoperability, which limits their effectiveness. This study proposes a CNN-BiLSTM-Attention (CBA) model for efficient attack detection using real-time UAV sensor data. Additionally, the SHapley Additive exPlanations (SHAP) method is used to improve the interpretability of the model. The proposed approach is tested on real attack scenarios, including denial-of-service (DoS) attacks and global positioning system (GPS) spoofing attacks, and demonstrates both effectiveness and interpretability.**



***Index Terms*— Attack detection, deep learning, unmanned aerial vehicles (UAVs).**

I. Introduction

RECENTLY, unmanned aerial vehicles (UAVs) have been widely used to perform various tasks in place of humans because of their high degree of flexibility. For example, they can act as aerial unmanned base stations and provide emergency Internet and communication services in the event of a disaster [1], [2]. However, many safety issues of UAVs are gradually exposed as UAVs become more popular, especially in cyber security. For example, Iran's capture of the American RQ-170 military UAV in 2011 was the most severe and far-reaching. From the perspective of the three elements of information security, these attacks mainly destroy the confidentiality, integrity, and availability of UAVs [3]. Technically, modern UAVs are vulnerable to denial of service (DoS), jamming, command injection, global positioning system (GPS) spoofing [4], [5], and other kinds of attacks. DoS and GPS spoofing attacks are simple and low-cost representative attacks. Specifically, DoS attacks disconnect commercial and civilian UAVs from ground control stations with communications congestion, and GPS spoofing deceives UAVs by forging unencrypted satellite signals [6]. Even though some excellent metaheuristics algorithms [7], [8], [9], [10], [11], [12], [13], [14] have started to be used in UAV path planning, both attacks can still quickly destabilize UAVs and, in the most severe cases, lead to the UAV crash or being captured by attackers. Therefore, it is necessary to detect them in a timely manner before they are deployed. While several methods are available [15], [16], [17], most are only suitable for specific types of attacks due to the attack specificity. For example, GPS spoofing attacks involve the manipulation of GPS signals to mislead individuals or devices. Detection methods for these attacks typically focus on identifying discrepancies between real and fake signals. However, these methods are limited in their ability to detect other types of attacks, such as DoS attacks, as they are solely based on GPS signals.

In this article, the author proposes a detection framework that uses SHapley Additive exPlanations (SHAP) to increase interpretability and is based on the generic state changes of UAVs using easily available status data from sensors such as GPS, inertial measurement unit (IMU), and gyroscope. Fig. 1 demonstrates how we use SHAP for local and global explanations of our model's decisions. In addition, neural network models are often used in attack detection, but their complexity can lead to a decrease in the interpretability and trustworthiness of the detection results, making them impractical. However, past studies have shown that the use of SHAP can provide interpretation and analysis for various models, including COVID-19 detection [18], power load forecasting [19], real-time accident detection [20], PM2.5 prediction [21], and intrusion detection systems (IDS) [22].

Manuscript received 4 January 2023; accepted 30 January 2023. Date of publication 17 February 2023; date of current version 31 March 2023. This work was supported in part by the Youth Program of the National Natural Science Foundation of China under Grant 62103330, and in part by the Key Program of the National Natural Science Foundation of China under Grant 62233014. The associate editor coordinating the review of this article and approving it for publication was Prof. You Li. *(Corresponding author: Yang Li.)*

Shihao Wu, Yang Li, Zheng Tan, and Quan Pan are with the School of Automation, Northwestern Polytechnical University, Xi'an 710129, China (e-mail: wshnpu@mail.nwpu.edu.cn; liyangnpu@nwpu.edu.cn; tanz@mail.nwpu.edu.cn; quanpan@nwpu.edu.cn).

Zhaoxuan Wang is with the School of Cybersecurity, Northwestern Polytechnical University, Xi'an 710129, China (e-mail: zxwang@mail.nwpu.edu.cn).

Digital Object Identifier 10.1109/JSEN.2023.3244831

Therefore, in this article, the authors propose a detection framework that uses SHAP to increase interpretability and is based on the generic state changes of UAVs using easily available status data from sensors such as the GPS, IMU, and gyroscope. Fig. 1 demonstrates how to use SHAP for local and global explanations of the model's decisions.

The local explanation explains why the model makes the final decision for each input. The global explanation shows the important features extracted from the model and the relationship between the feature values and different types of attacks. In this article, the authors address the challenge of maintaining high detection efficiency for small commercial and civil UAVs with limited computing resources. Then they examine the relationship between the physical characteristics of features and different types of attacks and identify the most important features for the analysis. To make efficient detection, they propose a hybrid model combining convolutional neural networks (CNNs) [23] and bi-directional long short-term memory (BiLSTM) [24] networks, with an attention mechanism, to enhance the performance of the baseline model. The experimental results demonstrate the effectiveness of this approach. The main contributions of this article are as follows.

1) This work is unique in the UAV security field. The SHAP method is used for the first time to improve the interpretability of the UAV attack detection model, which can help UAV security experts better understand the model's judgment and design the detection model's structure.
2) The authors revealed the relationship between UAV sensor data and different types of attacks in the actual physical meaning based on local and global explanations, explored the most effective sensor features for attack detection in the shortest time, and validated the minimum number of features required.
3) A CNN-BiLSTM-Attention (CBA) model that integrates the spatial features and temporal correlations of UAV sensor data is proposed, which achieved excellent and stable performance with limited computing resources of small commercial and civilian UAVs.

The rest of this article is organized as follows: Section II shows the related works on UAV attack detection. Section III gives a detailed description of the overall methodology. The dataset, experiments, and results are described in Section IV. Conclusion and future work are presented in Section V.

II. Related Works

With the increasingly widespread use of UAVs and the growing security concerns, numerous researchers have realized the importance of UAV security and conducted extensive research in UAV attack detection.

A. DoS Attack Detection

Chen et al. [25] presented a software framework that offers DoS attack-resilient control for real-time UAV systems using containers: ContainerDrone. A security monitor constantly checks DoS attacks over communication channels by simulating sensors and drivers in the container. The framework switches to the safety controller to mitigate the attack upon detecting a security rule violation. They implemented a prototype quadcopter with commercially off-the-shelf (COTS) hardware and open-source software. The experimental results demonstrated the effectiveness of the proposed framework in defending against various DoS attacks.

In addition to detecting DoS attacks from the software level, da-Silva et al. [26] proposed the development of an efficient platform based on the message queuing telemetry transport (MQTT) protocol for UAV control and denial-of- service (DoS) detection embedded in the UAV system. In DoS detection, the best results were a true positive rate (TPR) of 0.97 with 16 features from the AWID2 dataset using LightGBM with Bayesian optimization and data balancing. Unlike other studies, the built platform shows efficiency for UAV control and guarantees security in the communication with the broker and the Wi-Fi UAV network.

Furthermore, some researchers have tried to use neural networks to detect DoS attacks. Khan et al. [27] aimed to address the security deficiency by proposing an experience-based deep-learning algorithm to cater to the DoS attacks. The proposed scheme uses the IDS. The proposed approach is implemented as a case study in an innovative city environment. The result authenticates the superiority of the proposed schemes in terms of security and quality-of-service (QoS) requirements from their counterparts. Baig et al. [28] suggested a machine-learning-based approach for detecting hijacking, GPS signal jamming, and DoS attacks that can be carried out against a UAV. A detailed machine-learning-based classification of UAV datasets for the Da-Jiang innovations (DJI) Phantom 4 model was conducted, compromising both normal and malicious signatures. Results obtained yield advisory to foster futuristic opportunities to safeguard a UAV system against cyber-threats. Besides, supervised learning algorithms can protect UAVs from DoS attacks in multi-UAV systems by learning anomalous states [29].

B. GPS Spoofing Attack Detection

Panice et al. [30] first analyzed state estimation and then developed a support vector machine (SVM) as an anomaly detection tool for detection scheme and simulation environment for GPS spoofing attacks, which can be used to evaluate the functionality and performance of one class SVM.

In addition to detecting attacks by estimating the state of the UAV, Qiao et al. [31] proposed a vision sensor-based detection method to solve the GPS spoofing problem of small UAVs. They can detect spoofing attacks with an average of 5 s by using the UAV's sensors, monocular camera, and IMU to obtain the speed of the UAV.

There are also some methods for attack detection by extracting GPS signal features. Shafiee et al. [32] proposed a multitarget detection method based on multilayer neural network inputs, which performs spoofing detection by extending the traditional machine-learning algorithm k-nearest neighbors (KNN) and a simple Bayesian classifier using three main features extracted, that is, early–late phase, delta, and signal levels. Similarly, Manesh et al. [33] proposed a supervised

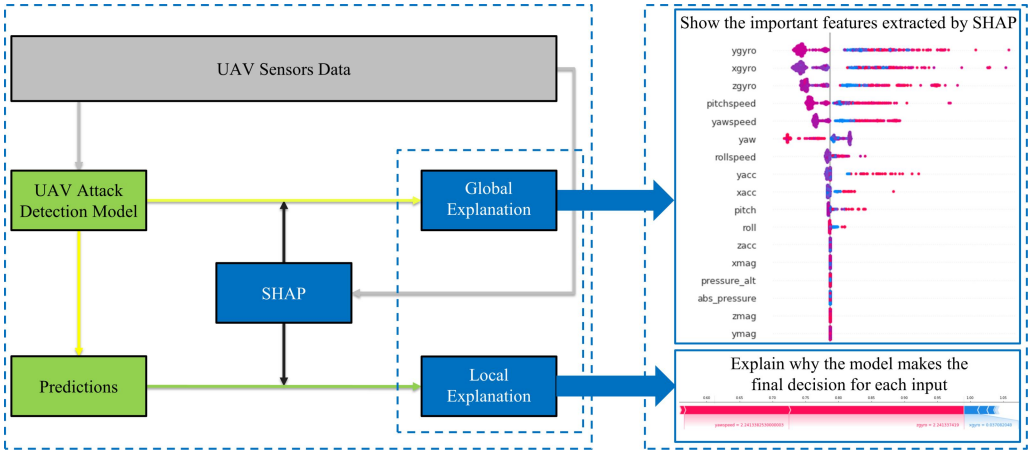


Fig. 1. General framework for local and global explanations of UAV attack detection models.

machine-learning method based on artificial neural networks to detect GPS spoofing signals. Different features such as pseudo-range, Doppler shift, and signal-to-noise ratio (SNR) are used to classify GPS signals. The results showed that the method has a high detection rate and a low false alarm probability. Meng et al. [15] proposed a spoof detection algorithm for UAV sensors based on GPS and optical flow fusion. The idea of the algorithm is to compare the fusion model of the raw GPS and optical flow data with the fusion data model of the UAV in its normal state and use the difference between them to determine whether the sensor is under attack.

Except for traditional machine-learning methods, deep-learning-based attack detection methods have also been applied. Xue et al. [16] proposed a deep-learning-based satellite image matching method, DeepSIM, for UAV GPS spoofing attacks. This method aims to compare historical satellite images of its GPS-based location with real-time aerial photos from its camera comparison. Practical experimental results show a success rate of about 95% in detecting GPS spoofing attacks within 100 ms. Kim et al. [17] used data augmentation to detect sensor spoofing. They developed a feed-forward depth model that captures the dynamic features of UAVs and generative adversarial networks (GANs) to augment the dataset for better training. The results show that Vanilla GANs are best suited for this task.

Unlike previous work, the authors used the SHAP method to explain UAV attack detection for the first time and proposed a framework that can effectively detect different types of UAV attacks. In addition, a hybrid model CBA based on global interpretation is proposed to select important features, which solves the problem of significant performance decrease of the baseline models after feature reduction and achieves excellent and stable attack detection performance with limited computational resources of small commercial and civilian UAVs.

III. Material and Methods

For small UAVs with limited computing resources, there is a need to detect attacks at the lowest possible cost. There are two main approaches to achieving this goal: reducing the number of parameters in the model or using fewer and more efficient data features. However, the first approach usually entails performance loss and instability. Therefore, this article focuses on how to use the most practical features to detect attacks. First, the SHAP method is used to make global and local explanations of the final decision of the UAV attack detection model, and important features are screened out based on the global explanations. However, during the experiments, the authors found that the performance of the baseline models significantly degrades after feature reduc- tion. To solve this problem, the authors propose a hybrid CNN-BiLSTM-attention (CBA) model to obtain better and more stable detection performance. The overall structure of the CBA model is shown in Fig. 2, which uses CNNs to extract spatial features and then sends the features extracted by CNNs to the BiLSTM network for processing and solv- ing the long-term data dependency problem. The attention mechanism can focus on the output features that are highly correlated with the detection results. In this section, the SHAP method and the components of the CBA model are presented separately.

A. Shapely Additive Explanations (SHAP)

SHAP is an additive attribution method proposed by Lundberg and Lee [34] to explain predictions based on Shapley values. The Shapley value is the unique solution in game theory that satisfies efficiency, symmetry, virtuality, and additivity. The game theory requires at least two things, a game and some players, and what Shapley does is quantify each player's contribution to the game. For SHAP, assuming that there is a prediction model, the model's prediction is the "game," and the samples contained in the model are the "players." The role of SHAP is to quantify the contribution of each feature to the prediction made by the model. In addition to the excellent properties of Shapley values, SHAP also has the desirable properties of local accuracy, missingness, and consistency. SHAP explains the predicted value of the model as the sum

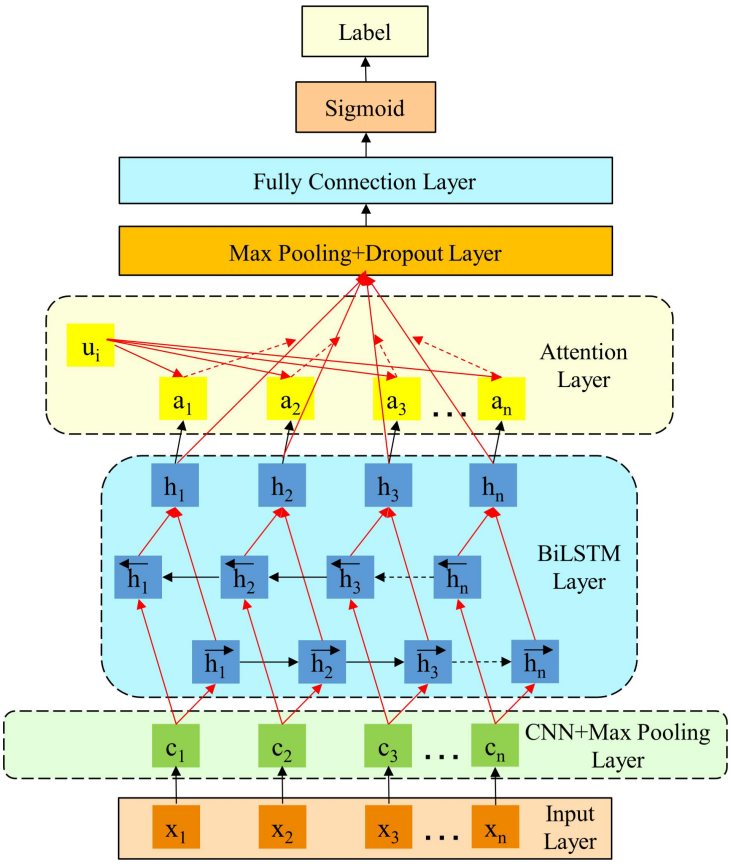


Fig. 2. Proposed hybrid network architecture.

of the attribute values of each input feature

$$g(z') = \varphi_0 + \sum_{j=1}^{M} \varphi_j z'_j \tag{1}$$

where g is the explanatory model, $z' \in \{0, 1\}^M$ indicates whether the corresponding feature can be observed (1 or 0), M is the number of input features, $\varphi_i \in \mathbb{R}$ is the attribute value (Shapley value) of each feature, and φ_0 is the constant of the explanatory model (i.e., the predicted mean of all training samples).

B. Convolutional Neural Network

Previous works [35], [36], [37], [38], [39], [40], [41] have demonstrated that CNNs outperform traditional machine-learning methods in many areas, such as computer vision and pattern recognition. The AlexNet network proposed by Krizhevsky et al. [23] introduced a new deep structure and dropout method that dramatically improved the accuracy of image recognition. Since then, CNNs have gained fame and flourished and are widely used in various fields, achieving the best current performance in many problems. In the CBA model, the input data is first transferred to the convolutional layer, which uses convolution kernels to convolve the input data to extract the spatial features. Since the sensor data of the UAV is sequence data, 1-D convolution (1-D-convolution) is applied here. The convolutional layer works as a filter and a nonlinear activation function. If the lth layer is a convolutional layer, the j th feature map of the lth layer is calculated as follows:

$$x_j^l = f\left(\sum_{i \in M_j} x_i^{l-1} \otimes k_{ij}^l + b_j^l\right) \tag{2}$$

where x_j^l is the corresponding activation and $\otimes$ is for the convolution kernel k to convolve all the feature maps associated with the $l-1$th layer and then sum them up and add a bias value b_j^l. f is a nonlinear function, the rectified linear unit (ReLU). For the 1-D-convolution setup, the number of filters is 16, and the kernel size is 4. As with the classical CNN, since 1-D-convolution will occupy the most weight parameters like the fully connected layer, the authors use a max pooling layer immediately after, which will optimize and reduce the model parameters to save memory and computational cost.

C. Bidirectional LSTM Network

The output features of the CNN will later be transferred to the BiLSTM network to solve the long-term data dependency problem. BiLSTM network is an excellent variant of the long short-term memory (LSTM) network, which consists of a forward LSTM network and a backward LSTM network, where the hidden layer at each moment in the LSTM network

contains multiple memory blocks. Each block has a cell (consisting of numerous memory cells) and three gates, that is, an input gate, a forgetting gate, and an output gate. The overall framework is shown in Fig. 3. The LSTM network first determines the information that needs to be discarded in the cell state, and this part of the operation is implemented by the sigmoid cell of the forgetting gate

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{3}$$

h_{t-1} denotes the activation at time $t-1$ and x_t is the input data at time t. The forgetting gate first splices h_{t-1} and x_t to obtain a long input vector and then multiplies the input vector by the weight matrix W_f to perform a fully connected computation, after which the bias term b_f is added to the result to obtain the hidden vector. Finally, the hidden vector is processed by the nonlinear activation function σ (i.e., sigmoid) to obtain the forgetting factor f_t, which is a vector between 0 and 1. The value in this vector indicates which information in the cell state C_{t-1} is retained or discarded, with 0 indicating no retention and 1 indicating all retention.

After processing the information that needs to be discarded, the next step is to determine what new information to add to the cell state, an operation that is divided into two stages. First, h_{t-1} and x_t are spliced, after which it is added with a bias value b_i and then an input gate is passed to determine which information to update. In addition, the new candidate cell information $\tilde{C}_t$ is obtained by passing the spliced h_{t-1} and x_t through a tanh function. The two steps are described as follows:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{4}$$

$$\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c). \tag{5}$$

The old cell information C_{t-1} will be updated to the new cell information C_t after deciding the further information to be added. The update rule is to select a part of the old cell information to be forgotten by the forget gate and a part of the candidate cell information $\tilde{C}_t$ to be added by the input gate to get the new cell information C_t. The update operation is described as follows:

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t. \tag{6}$$

After updating the cell information, the output gate needs to determine which state features of the cell need to be output based on the input h_{t-1} and x_t. The output gate first passes the spliced h_{t-1} and x_t through the sigmoid activation function to get the judgment condition, then passes the cell state information C_t through the tanh function to get a vector in the range of [−1, 1], and finally multiplies the vector with the obtained judgment condition to get the output. This step is described as follows:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{7}$$

$$h_t = o_t \cdot \tanh(C_t). \tag{8}$$

D. Attention Mechanism

To focus on the output features that are highly correlated with the detection results, the authors use the attention mechanism to calculate the weights of the feature vectors output by the BiLSTM network layer and flip the dimension using a permutation function. The attention mechanism will focus on the input data and assign different weights to the elements of the input sequence depending on the position and content of the sequence data

$$u_i = \tanh(Wh_i + b) \tag{9}$$

$$\alpha_i = \frac{\exp(u_i^T u_w)}{\sum_i \exp(u_i^T u_w)} \tag{10}$$

$$s = \sum_i \alpha_i h_i \tag{11}$$

where h_i denotes the hidden layer vector containing bidirectional sequence features at the output of the BiLSTM network. The attention mechanism first converts h_i to u_i through the fully connected layer (W is the weight vector of the attention mechanism and b is the bias value), then calculates the similarity between u_i and u_w, obtains the normalized weight vector α_i through the softmax function, and finally uses α_i as the weight to weight and sum h_i to obtain the output.

IV. Experiments

The authors first present the sensor data acquired in real UAV attack scenarios and the preprocessing process in this section. Then the authors construct a low-cost UAV attack detection framework for different attacks with high interpretability by answering the following four questions.

1) Analyze which baseline model performs best for attack detection.
2) Reveal which features data are most effective for attack detection.
3) Explore the minimum number of features required to ensure effective detection of an attack.
4) Find out how effectively the authors can detect an attack when a UAV is flying.

These four questions are a step-by-step relationship. In Question 1, the authors experimentally compared the performance of baseline models. The authors found the most suitable baseline model for UAV attack detection, which is the basis of our proposed CBA hybrid model. In Question 2, the authors preliminarily screened out the most valuable features for UAV attacks based on the global explanations made by the SHAP method. In Question 3, the authors specifically selected the minimum number of features required to effectively detect an attack based on the actual physical meaning of the features. The authors experimentally demonstrated the effectiveness of our screening strategy. In Question 4, the authors addressed the problem that the performance of the baseline models decreased significantly after feature reduction and proposed a CBA hybrid model that achieves excellent and stable performance with limited computational resources.

A. Dataset

To verify the effectiveness of our algorithm in real UAV attack scenarios, the authors performed a DoS and a GPS spoofing attack on the UAV shown in Fig. 4(a) by using the wireless network card in Fig. 4(b) and the Hack-RF one in Fig. 4(c), respectively. The DoS attack causes the UAV to lose

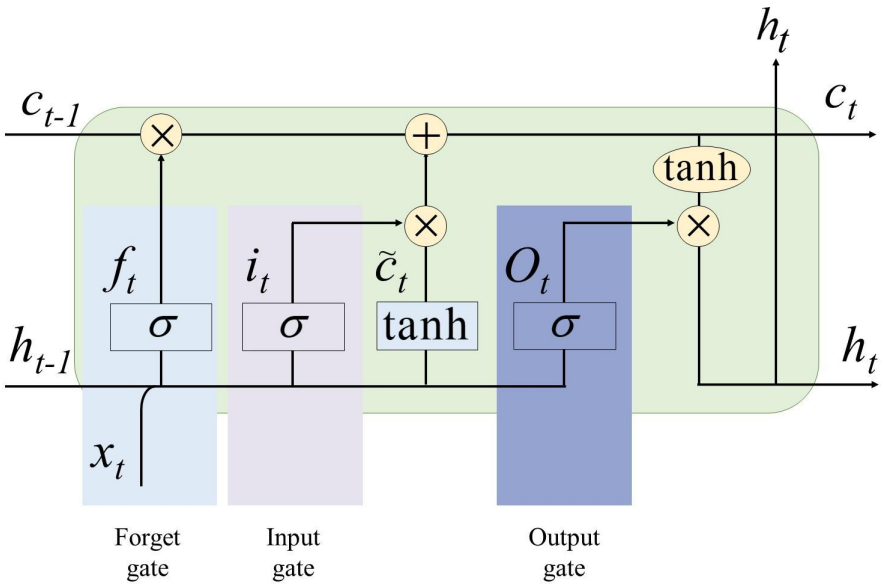


Fig. 3. Architecture of an LSTM block.

connection with the ground station and thus forces the UAV to land, while the GPS spoofing attack will force the UAV to fly to another wrong location. At the end of the flight, the authors use the UAV log files generated during the benign flight and the DoS and GPS spoofing attacks as our initial dataset. The initial dataset is ULog files containing the complete flight data.

B. Data Preprocessing

Algorithm 1 shows the whole data preprocessing process to get the initial dataset from the ULog files. The authors first converted the ULog files to comma-separated values (CSV) files via the ulog2csv script to facilitate reading the sensor data. Second, the sensor data scattered in different CSV files were concatenated by timestamp matching and labeled according to the attack time, solving the problem of many redundant records. After that, because the missing data is very slight, the authors removed the missing values to ensure the data's authenticity. To solve the problem of unbalanced data distribution due to the short attack time, the authors adopted a sampling-interpolation process: downsampling the unattacked data (i.e., the category with more data) and then the attacked data (i.e., the type with less data) was interpolated, that is, the average of every two adjacent data was taken to create a new data and inserted between the contiguous data, the comparison of results of sampling-interpolation processing are shown in Table I.

Similarly, to ensure the authenticity of the data, the authors kept only the interpolated data generated by this strategy in the training set. Still, the authors removed them from the test set. After that, the authors normalized the data with the following formula:

$$x_{\text{norm}} = (x - x_{\min})/(x_{\max} - x_{\min}). \tag{12}$$

The authors try to match the sensor data extracted from the CSV files by timestamps. Still, the complete set of sensor data that can be obtained under different types of attacks is slightly different due to the missing timestamps of some data. These sets, which will serve as the initial dataset for our subsequent experiments, are from the most important sensors on the UAV, such as the GPS, gyroscope, magnetometer, and accelerometer. The authors list them separately by attack type in Table II.

Algorithm 1 Process of Data Processing

```
Input: ULog file.
initialization;
Convert ULog into CSV by ulog2csv script;
for CSV from ULog do
    for sensors data in CSV do
        Merge the scattered sensors data by timestamp;
        Label the merged data by the attack time;
        Drop the labeled data with NaN values;
    end
end
for labeled data do
    for unattacked data do
        Downsampling the unattacked data;
    end
    for attacked data do
        Interpolating the attacked data;
    end
end
Divide the processed data into train set and test set;
for test set do
    Drop the data generated by interpolation;
end
Normalize the train set and the test set;
Output: Processed dataset for subsequent experiments.
```

C. QA1: Which Model Performs Best?

Since UAV data are sequence data with certain spatial features, the authors supplemented traditional machine-learning

(a) (b) (c)

Fig. 4. Equipment used in the attack experiments. (a) UAV. (b) Wireless card. (c) HackRF one.

TABLE I
COMPARISON OF SAMPLING-INTERPOLATION PROCESSING RESULTS

Comparison	The ratios of unattacked data to attacked data
Before sampling-interpolation processing	5:1
After sampling-interpolation processing	1:1

TABLE II
INITIAL DATASET AFTER PROCESSING UNDER DIFFERENT ATTACKS

Source Sensor	Data	Data Type	Data Description	Data Range	Attack Type
Accelerometer	xacc	float	X-axis acceleration	[0,1]	DoS Attack GPS Spoofing Attack
	yacc	float	Y-axis acceleration	[0,1]	
	zacc	float	Z-axis acceleration	[0,1]	
Magnetometer	xmag	float	X-axis geomagnetic	[0,1]	
	ymag	float	Y-axis geomagnetic	[0,1]	
	zmag	float	Z-axis geomagnetic	[0,1]	
Gyroscope	xgyro	float	X-axis angular velocity	[0,1]	
	ygyro	float	Y-axis angular velocity	[0,1]	
	zgyro	float	Z-axis angular velocity	[0,1]	
Inertial measurement unit (IMU)	pitch	float	Pitch angle	[0,1]	DoS Attack
	yaw	float	Yaw angle	[0,1]	
	roll	float	Roll angle	[0,1]	
	pitchspeed	float	Pitch speed	[0,1]	
	yawspeed	float	Yaw speed	[0,1]	
	rollspeed	float	Roll speed	[0,1]	
Barometer	pressure_alt	float	Altitude	[0,1]	
	abs_pressure	float	Absolute value of altitude	[0,1]	
GPS	x	float	X-axis coordinate	[0,1]	GPS Spoofing Attack
	y	float	Y-axis coordinate	[0,1]	
	z	float	Z-axis coordinate	[0,1]	
	vx	float	X-axis velocity	[0,1]	
	vy	float	Y-axis velocity	[0,1]	
	vz	float	Z-axis velocity	[0,1]	
/	label	int	Data label	0 or 1	/

algorithms SVM-RBF and back-propagation (BP) neural network for comparison to demonstrate the effectiveness of CNNs in extracting spatial features and the ability of LSTM to solve the long-term dependence of sequence data.

Except for SVM-RBF, the three neural network models mentioned above were all built using Keras in the TensorFlow backend. Since our purpose is to compare the suitability of different models for UAV attack detection, all models use a simple standard structure with the same order of magnitude parameters. ReLU and Sigmoid activation functions are used for the models' intermediate hidden and output layers. All models use the Adam optimizer and are trained for 150 epochs. These four models' loss curves and detection results for the two attacks are shown in Fig. 5 and Table III, respectively.

For DoS attack detection, the CNN performs best and can converge quickly during training. LSTM and SVM have about the same performance. Still, LSTM is challenging to converge rapidly at the beginning of training, BP has the worst performance, and the loss value when reaching convergence is challenging to get the level of CNN and LSTM. For GPS spoofing attack detection, the performance and convergence speed of BP, CNN, and LSTM are about the same, with LSTM having the best detection performance and SVM the worst. The authors believe that the difference in convergence speed in training is that a GPS spoofing attack generally causes UAVs to fly to the wrong location. In contrast, a DoS attack causes UAVs to disconnect, stop flying or even crash. The corresponding sensors' data will change more drastically than when subjected to GPS spoofing attacks, so it is more difficult for the models to learn valuable features. In summary, both CNN and LSTM can effectively handle UAV sensor data and achieve reasonable attack detection rates. At the same time,

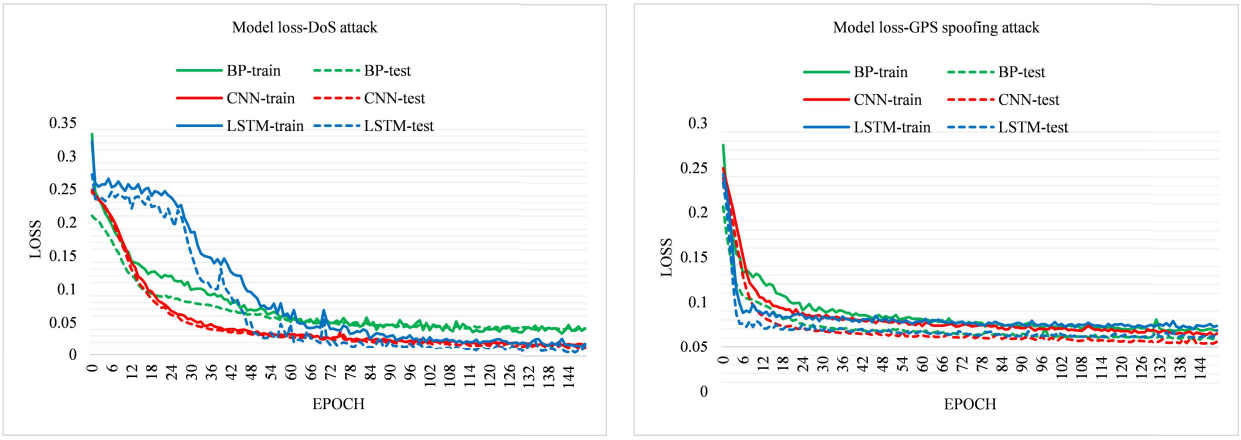


Fig. 5. Loss curves of baseline models. (a) DoS attack detection. (b) GPS spoofing attack detection.

TABLE III
DETECTION RATES OF FOUR BASELINE MODELS FOR TWO TYPES OF ATTACKS

Model	CNN	LSTM	BP	SVM
DoS attack	93.6%	94.7%	91.2%	94.7%
GPS spoofing attack	92.8%	91.6%	91.9%	90.6%

BP and SVM-RBF are unstable. So, the authors chose CNN and LSTM as the essential components of our hybrid CNN-BiLSTM-Attention model.

D. QA2: Which Types of Data Are Most Effective?

Since small commercial and civilian UAVs have minimal computational resources and attack detection requires high timeliness, it is necessary to find the most efficient UAV sensor data to save resources and improve the timeliness of detection. Moreover, as sophisticated neural network models are used in attack detection, there is a decrease in the interpretability of the detection results as the complexity of the model increases, which makes the untrustworthiness of the detection results leading to impracticality. However, the role of sensor data in detecting different types of attacks should be consistent with its actual physical meaning. To explain this issue and enhance the interpretability of attack detection, the authors used SHAP to interpret the model's decisions globally. On this basis, the authors explained the relationship between the actual physical meaning of the features and different types of attacks.

Fig. 6 shows the global explanations of the model's decision-making of two different attack detections, and all features extracted by the DoS attack and GPS spoofing attack are sorted from highest to lowest importance. Each point in Fig. 6 represents feature data, the y-axis represents the global importance of different features in model decision-making, in order from highest to lowest, and the x -axis represents the Shapely value of different feature data, which means the degree of contribution to the model decision, linearly related to the degree of contribution. The color indicates the feature values from low to high, with the intensity of red increasing as the feature values increase and the power of blue increasing as the feature values decrease.

There are significant differences in the important features extracted under the two attacks. In a DoS attack, the most important features are from the IMU, gyroscope, and accelerometer. In a GPS spoofing attack, the most important features are mainly from GPS rather than IMU or gyroscope. The authors believe the reason for such a significant difference is that DoS attacks and GPS spoofing attacks have different effects on UAVs. In most cases, a DoS attack causes the UAV to disconnect, stop flying, or even crash, thus causing dramatic changes in sensor data. In contrast, a GPS spoofing attack spoofs the UAV to another location where the UAV still usually flies, and accordingly, GPS data changes significantly.

E. QA3: The Minimum Number of Features Needed

Since the DoS attack and GPS spoofing attack can cause the UAV to disconnect, stop flying, crash, or fly to another location, the sensor data of the UAV during the attack is disordered and will only change drastically without specific rules. Positive and negative, too-large, and too-small feature values are likely to play an equally important role in model decisions, so the authors cannot observe a linear relationship between SHAP values and feature values in the global explanations. Even so, because the sensor data the authors used are sequence data, the feature values will have a gradual process on the time axis, so points of similar color in the global interpretations will be clustered together on the x -axis. For these reasons, although the authors obtained the most important features under particular attacks based on the global explanations of the model decisions, and these features also play an important role in detecting such attacks, the authors still cannot easily determine the minimum number of features required for attack detection.

From the global explanations shown in Fig. 6, the data from IMU and gyroscope play a more important role in DoS attack detection, while the data from the accelerometer on the x -axis and y-axis are slightly less important, and the data on the z-axis are even ranked last. In GPS spoofing attack detection, the data that plays a more important role comes from the GPS,

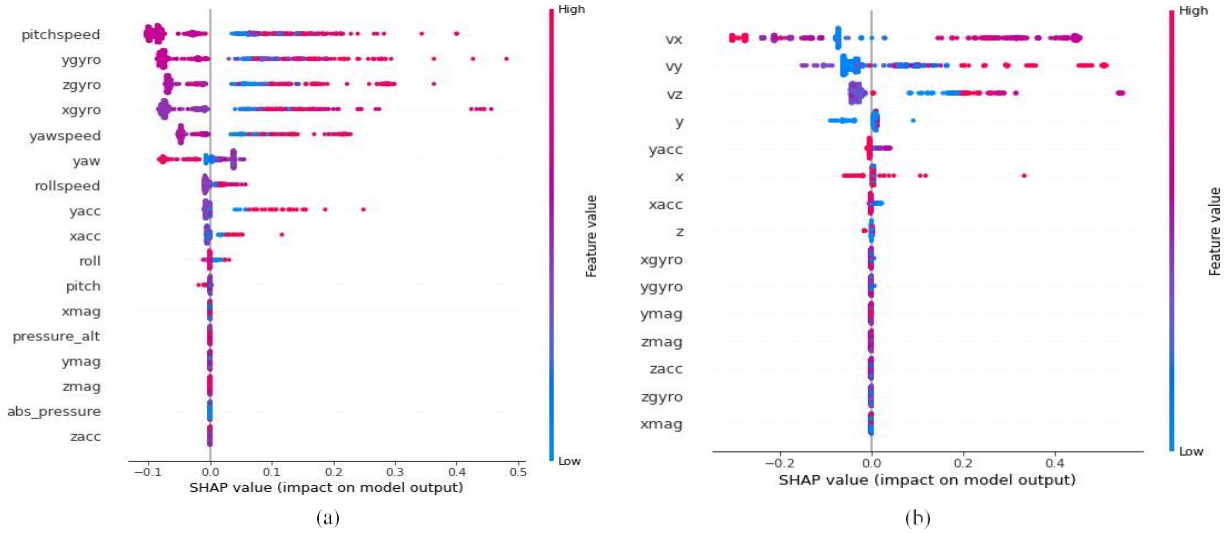


Fig. 6. Global explanations of DoS attack and GPS spoofing attack detections. (a) Global explanation of DoS attack detection. (b) Global explanation of GPS spoofing attack detection.

TABLE IV
DETECTION RATES OF TWO ATTACKS AFTER FEATURE REDUCTION

	DoS attack			GPS spoofing attack
Model	IMU data	Gyroscope data	IMU and Gyroscope data	GPS data
CNN	81.7%	83.2%	95.4%	92.1%
LSTM	88.4%	87.1%	93.6%	91.4%
BP	84.2%	81.6%	92.9%	96.2%
SVM	80.4%	79%	93.2%	88.2%

and the data from the accelerometer has the same problem as the data in the DoS attack.

Considering the physical meaning of the features and the practical application significance of the attack detection, the shorter the delay in acquiring the data, the lower the cost of attack detection if there are fewer source sensors to obtain the data. To this end, the authors used the following strategy in our experiments to explore the minimum number of features required: in DoS attack detection, the authors use only features from the IMU and gyroscope and compare them with features from the IMU or gyroscope only, and in GPS spoofing attack detection, the authors use only features from GPS. The experimental results according to the above strategy are shown in Table IV.

Although the reduced number of features used slightly reduces the detection rate for both attacks, it still justifies our strategy above. Using data from a gyroscope or IMU alone causes a significant detection error for the DoS attack. Using a mixture of the gyroscope and IMU causes a slight loss in detection rate. The minimum number of features required under both attack detections are detailed in Table V.

TABLE V
MINIMUM NUMBER OF FEATURES REQUIRED FOR DIFFERENT TYPES OF ATTACK DETECTIONS

Source Sensor	The features used	Attack Type
Gyroscope	xgyro	DoS Attack
	ygyro	
	zgyro	
Inertial measurement unit (IMU)	pitch	
	yaw	
	roll	
	pitchspeed	
	yawspeed	
	rollspeed	
GPS	x	GPS Spoofing Attack
	y	
	z	
	vx	
	vy	
	vz	

F. QA4: How Effective Can the Authors Detect?

In determining the minimum number of features required to detect an attack, although the detection time is less, the performance of all baseline models decreases due to the reduction of features, which is a fatal problem for UAV attack detection. To combine the detection rate and timeliness, the authors use our CBA hybrid model to detect attacks effectively. The main hidden layer of CBA consists of a convolutional layer, a BiLSTM layer, and an attention layer, ensuring that it has the same parameters and detection time as the baseline models. Table VI shows the results in detail.

CBA compensates for the loss of detection rate due to feature reduction and achieves a better detection rate. The number of parameters and detection time of CBA remain in the same order of magnitude as the baseline models without increasing the cost and latency of UAV attack detection. In addition, the authors analyzed the convergence of CBA to ensure fairness. The authors compared the CBA with the baseline models before feature reduction because this did not

TABLE VI
COMPARISON OF THE CBA MODEL AND BASELINE MODELS

Model	Parameters	Detection time	DoS attack	GPS spoofing attack
CNN	1481	160us	95.4%	92.1%
LSTM	1697	692us	93.6%	91.4%
BP	689	54us	92.9%	96.2%
CBA	1956	367us	99.4%	99.1%

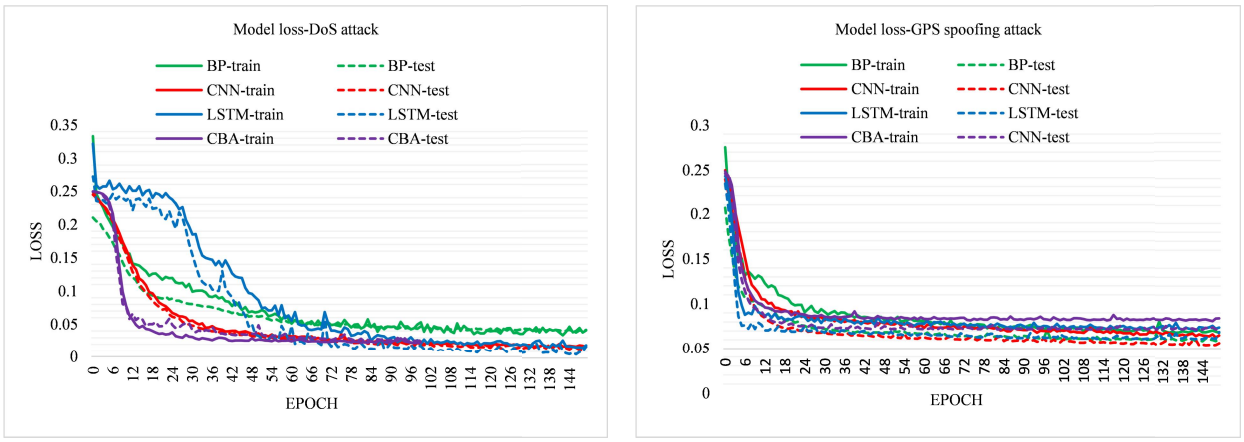


Fig. 7. Loss curves for the CBA model and baseline models. (a) DoS attack detection. (b) GPS spoofing attack detection.

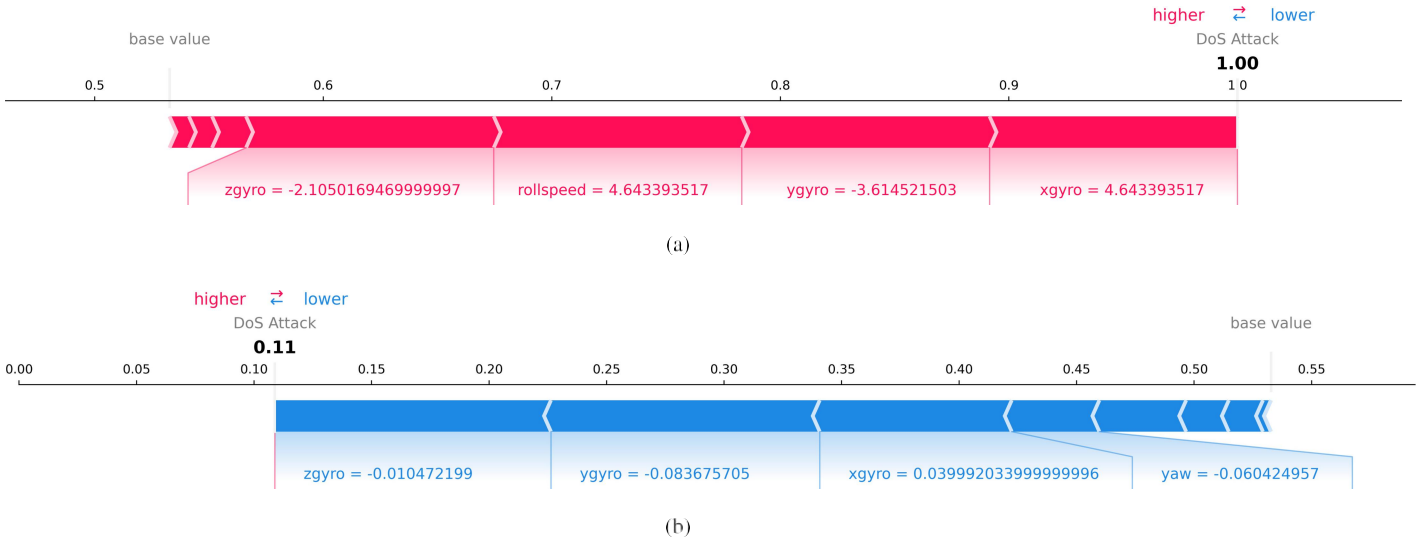


Fig. 8. Local explanations of DoS attack detection. (a) Local explanation of a positive sample. (b) Local explanation of a negative sample.

cause a significant decrease in the performance of the baseline models. As seen in Fig. 7, CBA converges faster than the baseline models in DoS attack detection, and there is almost no difference between CBA and the baseline models in GPS spoofing attack detection.

In the actual application environment, the sensor data of the UAV will continuously generate sequence data, and each new data will disrupt the whole data distribution, so it cannot be normalized in real-time. During the experiment, the authors found that some baseline models could not converge when using nonnormalized data. Still, the CBA model performed much better, which indicates that the CBA model can be more effectively adapted to the practical application environment.

In summary, CBA can improve the accuracy of UAV attack detection while ensuring timeliness and making the detection more efficient.

The authors use SHAP to explain decisions on individual data and randomly select two data samples for each of the two attacks. Fig. 8 shows the contribution of each feature value to DoS attack detection, and Fig. 9 shows the contribution of each feature value to GPS attack detection. The bold font indicates CBA's confidence that the data is attacked data, just as in the global explanations, red shows that the feature has a positive impact on the decision, and blue means that the feature harms the decision. It can also be seen that all features play a positive role in the decision where the CBA model

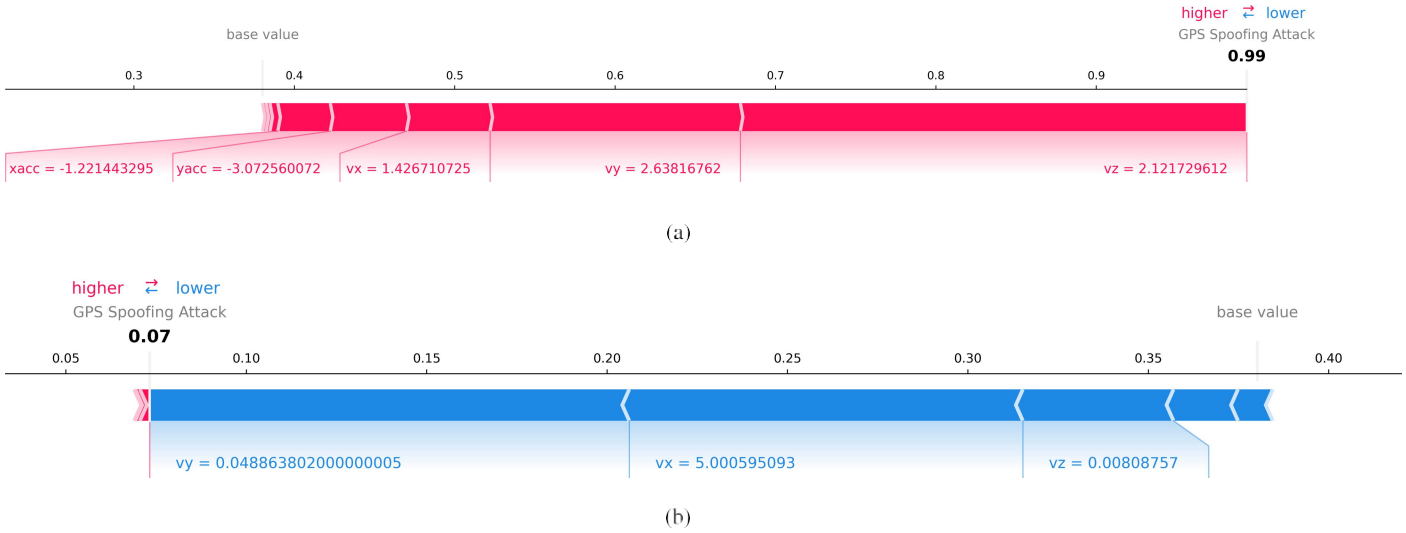


Fig. 9. Local explanations of GPS spoofing attack detection. (a) Local explanation of a positive sample. (b) Local explanation of a negative sample.

TABLE VII
RESULTS OF ABLATION EXPERIMENTS

Model	DoS attack	GPS spoofing attack
CNN	95.4%	92.1%
BiLSTM	88.2%	90.4%
CNN-BiLSTM	92.2%	92.8%
CNN-Attention	90.7%	87.6%
BiLSTM-attention	93.7%	95.4%
CBA	99.4%	99.1%

judges the data to be attacked data. All features play a negative role in the decision where the CBA model evaluates the data as unattacked data, indicating that the features the authors identified in Section IV-E are pretty accurate.

G. Ablation Study

To better understand CBA, the authors conducted ablation experiments to compare the performance of hybrid models composed of different components. Among all the results shown in Table VII, CBA had the highest detection rates for both attacks, reaching 99.4% and 99.1%, respectively. To our surprise, the performance of some other hybrid models is even lower than the baseline models. The results of the ablation experiments show that the effectiveness of CBA comes from the joint action of the components that can make it more effective in detecting different types of attacks, which can be explained by the following mechanisms: the CNN is used to extract spatial features. The features extracted by the CNN are fed to the BiLSTM network for processing. BiLSTM network is added to address the long-term dependency of the data, and the attention mechanism is used to focus on the BiLSTM network layer output features that are highly correlated with the detection results.

V. CONCLUSION AND FUTURE WORKS

In this article, the authors proposed a framework for UAV attack detection with higher interpretability based on UAV state information, which can effectively detect different types of UAV attacks. Starting from classical baseline models, the authors explored which baseline models are more suitable for drone attack detection and used the SHAP method to provide local and global explanations of the model's final decisions for the first time. The local explanations explain why the model makes the final decision for each input. The global explanations show the important features extracted from the model and the relationship between the feature values and different types of attacks. Based on the global explanations, the authors explained the relationship between the physical meaning of the features and the different types of attacks, screening out the features with lower contributions. Given the limited computational resources of small commercial and civilian UAVs, the authors determined the minimum number of features required for different attacks based on the global explanations and the physical meaning of the features. Since the performance of the baseline models decreased significantly after feature reduction, the authors selected the most suitable models for UAV attack detection. The authors proposed the hybrid model CBA, which can improve the accuracy of UAV attack detection while ensuring timeliness and making the detection more effective. Finally, the authors conducted DoS attacks and GPS spoofing attacks on UAVs in real scenarios, respectively, and validated the effectiveness of the proposed framework and hybrid CBA model by the acquired actual sensors data, and the results showed that our approach is superior to other methods.

Information security research on UAVs will be the focus of our future work, and interesting research directions include: 1) designing more methods for information security attacks on UAVs; 2) exploring more possibilities for detecting information security attacks on UAVs using sensors data fusion; and 3) studying the security of artificial intelligence algorithms deployed on UAVs.

REFERENCES

[1] A. Eldosouky, W. Saad, and N. Mandayam, "Resilient critical infrastructure: Bayesian network analysis and contract-based optimization," *Rel. Eng. Syst. Saf.*, vol. 205, Jan. 2021, Art. no. 107243.

[2] R. Amer, W. Saad, H. ElSawy, M. M. Butt, and N. Marchetti, "Caching to the sky: Performance analysis of cache-assisted CoMP for cellular-connected UAVs," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2019, pp. 1–6.

[3] D. He, X. Du, Y. Qiao, Y. Zhu, Q. Fan, and W. Luo, "A survey on cyber security of unmanned aerial vehicles," *Jisuanji Xuebao*, vol. 42, no. 5, pp. 1076–1094, Jul. 2019.

[4] K. Parlin, M. M. Alam, and Y. L. Moullec, "Jamming of UAV remote control systems using software defined radio," in *Proc. Int. Conf. Mil. Commun. Inf. Syst. (ICMCIS)*, May 2018, pp. 1–6.

[5] T. P. Vuong, G. Loukas, D. Gan, and A. Bezemskij, "Decision tree-based detection of denial of service and command injection attacks on robotic vehicles," in *Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, Nov. 2015, pp. 1–6.

[6] M. R. Manesh and N. Kaabouch, "Cyber-attacks on unmanned aerial system networks: Detection, countermeasure, and future research directions," *Comput. Secur.*, vol. 85, pp. 386–401, Aug. 2019.

[7] A. E. Ezugwu, J. O. Agushaka, L. Abualigah, S. Mirjalili, and H. A. Gandomi, "Prairie dog optimization algorithm," *Neural Comput. Appl.*, vol. 34, pp. 20017–20065, Jul. 2022.

[8] J. O. Agushaka, A. E. Ezugwu, and L. Abualigah, "Dwarf mongoose optimization algorithm," *Comput. Methods Appl. Mech. Eng.*, vol. 391, Mar. 2022, Art. no. 114570.

[9] L. Abualigah, D. Yousri, M. A. Elaziz, A. A. Ewees, M. A. A. Al-Qaness, and A. H. Gandomi, "Aquila optimizer: A novel meta-heuristic optimization algorithm," *Comput. Ind. Eng.*, vol. 157, Jul. 2021, Art. no. 107250.

[10] L. Abualigah, M. A. Elaziz, P. Sumari, Z. W. Geem, and A. H. Gandomi, "Reptile search algorithm (RSA): A nature-inspired meta-heuristic optimizer," *Expert Syst. Appl.*, vol. 191, Apr. 2022, Art. no. 116158.

[11] O. N. Oyelade, A. E.-S. Ezugwu, T. I. A. Mohamed, and L. Abualigah, "Ebola optimization search algorithm: A new nature-inspired Metaheuristic optimization algorithm," *IEEE Access*, vol. 10, pp. 16150–16177, 2022.

[12] L. Abualigah, A. Diabat, S. Mirjalili, M. A. Elaziz, and A. H. Gandomi, "The arithmetic optimization algorithm," *Comput. Methods Appl. Mech. Eng.*, vol. 376, Apr. 2021, Art. no. 113609.

[13] L. Abualigah and A. Diabat, "Advances in sine cosine algorithm: A comprehensive survey," *Artif. Intell. Rev.*, vol. 54, pp. 2567–2608, Jan. 2021.

[14] L. Abualigah, "Feature selection and enhanced krill herd algorithm for text document clustering," in *Studies in Computational Intelligence*. Cham, Switzerland: Springer, 2019.

[15] L. Meng, S. Ren, G. Tang, C. Yang, and W. Yang, "UAV sensor spoofing detection algorithm based on GPS and optical flow fusion," in *Proc. 4th Int. Conf. Cryptography, Secur. Privacy*, Jan. 2020, pp. 146–151.

[16] N. Xue, L. Niu, X. Hong, Z. Li, L. Hoffaeller, and C. Pöpper, "DeepSIM: GPS spoofing detection on UAVs using satellite imagery matching," in *Proc. Annu. Comput. Secur. Appl. Conf.*, Dec. 2020, pp. 304–319.

[17] K. Kim et al., "Security analysis against spoofing attacks for distributed UAVs," in *Proc. ACM Conf Comput. Commun. Secur. (CCS)*, Vienna, Austria, Oct. 2016, pp. 1–6, doi: 10.1145/2976749.2978388.

[18] K. Debjit et al., "An improved machine-learning approach for COVID-19 prediction using Harris hawks optimization and feature analysis using SHAP," *Diagnostics*, vol. 12, no. 5, p. 1023, Apr. 2022.

[19] M. Li and Y. Wang, "Power load forecasting and interpretable models based on GS_XGBoost and SHAP," *J. Phys., Conf. Ser.*, vol. 2195, no. 1, Feb. 2022, Art. no. 012028.

[20] A. B. Parsa, A. Movahedi, H. Taghipour, S. Derrible, and A. Mohammadian, "Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis," *Accident Anal. Prevention*, vol. 136, Mar. 2020, Art. no. 105405.

[21] A. C. Just et al., "Advancing methodologies for applying machine learning and evaluating spatiotemporal models of fine particulate matter ($PM_{2.5}$) using satellite data over large regions," *Atmos. Environ.*, vol. 239, Oct. 2020, Art. no. 117649.

[22] T.-T.-H. Le, H. Kim, H. Kang, and H. Kim, "Classification and explanation for intrusion detection system based on ensemble trees and SHAP method," *Sensors*, vol. 22, no. 3, p. 1154, Feb. 2022.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[24] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," 2015, *arXiv:1508.01991*.

[25] J. Chen, Z. Feng, J.-Y. Wen, B. Liu, and L. Sha, "A container-based DoS attack-resilient control framework for real-time UAV systems," in *Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE)*, Mar. 2019, pp. 1222–1227.

[26] L. M. da Silva et al., "Development of an efficiency platform based on MQTT for UAV controlling and DoS attack detection," *Sensors*, vol. 22, no. 17, p. 6567, Aug. 2022.

[27] I. U. Khan, A. Abdollahi, M. A. Khan, I. Uddin, and I. Ullah, 'Securing against DoS/DDoS attacks in internet of flying things using experience-based deep learning algorithm," *Res. Square*, 2021, doi: 10.21203/rs.3.rs-271920/v1.

[28] Z. Baig, N. Syed, and N. Mohammad, "Securing the smart city airspace: Drone cyber attack detection through machine learning," *Future Internet*, vol. 14, no. 7, p. 205, Jun. 2022.

[29] C. Rani, H. Modares, R. Sriram, D. Mikulski, and F. L. Lewis, "Security of unmanned aerial vehicle systems against cyber-physical attacks," *J. Defense Modeling Simul.*, vol. 13, no. 3, pp. 331–342, Jul. 2016.

[30] G. Panice et al., "A SVM-based detection approach for GPS spoofing attacks to UAV," in *Proc. 23rd Int. Conf. Autom. Comput. (ICAC)*, Sep. 2017, pp. 1–11.

[31] Y. Qiao, Y. Zhang, and X. Du, "A vision-based GPS-spoofing detection method for small UAVs," in *Proc. 13th Int. Conf. Comput. Intell. Secur. (CIS)*, Dec. 2017, pp. 312–316.

[32] E. Shafiee, M. R. Mosavi, and M. Moazedi, "Detection of spoofing attack using machine learning based on multi-layer neural network in single-frequency GPS receivers," *J. Navig.*, vol. 71, no. 1, pp. 169–188, Jan. 2018.

[33] M. R. Manesh, J. Kenney, W. C. Hu, V. K. Devabhaktuni, and N. Kaabouch, "Detection of GPS spoofing attacks on unmanned aerial systems," in *Proc. 16th IEEE Annu. Consum. Commun. Netw. Conf. (CCNC)*, Jan. 2019, pp. 1–6.

[34] S. M. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2017, pp. 4766–4775.

[35] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent., ICLR Conf. Track*, San Diego, CA, USA, May 2015, pp. 1–14.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[37] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. CVPR*, Jul. 2017, pp. 1492–1500.

[38] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2261–2269.

[39] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Nov. 2018, pp. 7132–7141.

[40] A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[41] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jan. 2018, pp. 4510–4520.

Shihao Wu received the bachelor's degree from Northwestern Polytechnical University, Xi'an, China, in 2020, and the master's degree from the School of Cybersecurity, Northwestern Polytechnical University, in 2022, where he is pursuing the Ph.D. degree with the School of Automation.

His research interests include unmanned aerial vehicle (UAV) security, pattern recognition, adversarial attack, and explainable artificial intelligence.

Yang Li (Member, IEEE) received the bachelor's and Ph.D. degrees from Northwestern Polytechnical University, Xi'an, China, in 2014 and 2018, respectively.

He worked as a Research Fellow of the Sentic Team under Professor Erik Cambria at Nanyang Technological University, Singapore, and also was an Adjunct Research Fellow at the A*STAR High-Performance Computing Institute (IHPC), Singapore. He is an Associate Professor with the School of Automation, Northwestern Polytechnical University. He has published several papers on these topics at international conferences and in peer-reviewed journals. His research interests are in adversarial attack and defense in AI, NLP, recommender systems, and explainable artificial intelligence.

Dr. Li is an Active Reviewer of several journals, for example, INFORM FUSION, IEEE Transactions on Affective Computing (TAFF), neurocomputing (NEUCOM), knowledge-based system (KBS), and Knowledge and information system (KAIS). He is also an Advisory Board Member of Socio-Affective Computing and the Guest Editor of *Future Generation Computer Systems*.

Zhaoxuan Wang received the B.S. degree from the School of Cybersecurity, Northwestern Polytechnical University, Xi'an, China, in 2020, where he is pursuing the Ph.D. degree with the School of Cybersecurity.

His research interests include unmanned aerial system security, autonomous driving security, and artificial intelligence security.

Zheng Tan was born in China, in 1993. He received the master's degree from the Civil Aviation University of China, Tianjin, China, in 2020. He is currently pursuing the Ph.D. degree with Northwestern Polytechnical University, Xi'an, China.

His research interests include information fusion for robotic systems and UAV control and navigation.

Quan Pan (Member, IEEE) was born in China, in 1961. He received the B.S. degree in automatic control from the Huazhong University of Science and Technology, Wuhan, China, in 1982, and the M.S. and Ph.D. degrees in control theory and application from Northwestern Polytechnical University, Xi'an, China, in 1991 and 1997, respectively.

From 1991 to 1993, he was an Associate Professor with Northwestern Polytechnical University, where he has been a Professor with the Automatic Control Department, since 1997. He has authored 11 books and more than 400 articles. His research interests include information fusion, target tracking and recognition, deep network and machine learning, UAV detection navigation and security control, polarization spectral imaging and image processing, industrial control system information security, commercial password applications, and modern security technologies.

Dr. Pan is an Associate Editor of the journal *Information Fusion and Modern Weapons Testing Technology*.