Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Empirical Stability Assessment of SHAP-Based Feature Selection Across Controlled Model Complexities

Cynthia C

p20210415@hyderabad.bits-pilani.ac.in

BITS-Pilani, Hyderabad Campus

Gopal Krishna Kamath

IIT Dharwad

Debayani Ghosh

Thapar Institute of Engineering and Technology

Prashant K. Wali

BITS-Pilani, Hyderabad Campus

Article

Keywords: Machine Learning Mode, SHAP, Convexity, Submodularity

Posted Date: October 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11140613/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Empirical Stability Assessment of SHAP-Based Feature Selection Across Controlled Model Complexities

Cynthia C[1,2*], Gopal Krishna Kamath[3], Debayani Ghosh[4], Prashant K. Wali[1]

[1*]Electrical Engineering, BITS-Pilani, Hyderabad Campus, Jawahar Nagar, Secunderabad, 5000078, Telangana, India.
[2]AI and Data Science Engineering, Christ University, Kengeri Campus, Mysore Road, Bangalore, 560074, Karnataka, India.
[3]Electrical, Electronics and Communication Engineering, IIT Dharwad, Chikka Malligwad, Dharwad, 580011, Karnataka, India.
[4]Electronics and Communication Engineering, Thapar Institute of Engineering and Technology, Adarsh Nagar, Patiala, 147004, Punjab, India.

*Corresponding author(s). E-mail(s): p20210415@hyderabad.bits-pilani.ac.in;
Contributing authors: gopal.kamath@iitdh.ac.in; debayani.ghosh@thapar.edu; wali@hyderabad.bits-pilani.ac.in;

Abstract

Currently, the contemporary world is highly dependent on many automated decision-making systems driven by machine learning models. The models rely on statistical and computational methods, they learn patterns from the data to perform tasks such as prediction and classification. Understanding how the input features influence the model outputs, particularly as model complexity increases, is a critical aspect to be investigated. This study investigated the impact of model complexity on feature attribution and interpretability by assessing the tolerance and variability different models. The research employed the SHAPLEY approach, the primary analytical framework for evaluation. It examines how model complexity influences the stability, reliability and uncertainty of the feature importance estimates. To further comprehend the working of the SHAPLEY in association with the machine learning model, the model's optimization

complexity was identified. To observe the operation of the SHAPLEY, the custom model f(x) was introduced in the proposed work. Model f(x) has both linear and non-linear parts, which can be scaled by the factor alpha. However, a significant lacuna is evident in existing research models, which is related to the detailed analysis of the SHAPLEY-selected features over the black-box model and its variability measures. Our novel approach represents a unique way of assessing the SHAPLEY selected features using the custom model f(x). Additionally, the model complexity was identified and its influence on the SHAPLEY was performance. The performance of the proposed model was validated using the CICIDS2017 dataset.

Keywords: Machine Learning Mode, SHAP, Convexity, Submodularity

1 Introduction

In today's world, many critical decisions from medical, financial, and cybersecurity perspectives are being guided by machine learning models. These models mainly rely mainly on two factors: input features and model training using a selected set of features. As model complexity increases, understanding how decisions are influenced by selected input features and their attributions is essential.

Specifically, in the world of cybersecurity, the incorporation of machine learning to identify various cyber threats, specifically DDoS (Distributed Desnial-of-Service) attacks, has marked pivotal development. DDoS attacks are known for their ability to overload networks with a stream of web traffic and pose a significant challenge in developing good predictive models. The papers [1–9] highlight the critical role of machine learning models in detecting DDoS attacks.

The key to the model prediction accuracy depends on the diligent selection of features. However, selecting the best features is a nuanced process; the traditional feature selection techniques are mostly model-specific, and they give the overall feature importance for the collected data. Thus, the Shap[1] approach has been incorporated in the proposed framework, which is model-agnostic, to identify the best features for model prediction.

Broadly, there are two common workflows: first, the models make predictions and methods such as SHAP are used to explain the contribution of each feature and the other is where the SHAP based attributions are used to guide the feature selection itself in the initial phase to refine the model and the predictions are made followed by the process again.

SHAP, grounded in game theory, provide the explanation, importance, and contribution of influence of each feature on the decision made. It is based on SHAP values tagged to each feature, these values are supposed to be create fairness. The application of the SHAP-based approach for the purpose of selecting the best feature based on model performance is an emerging trend in recent research, as shown in [4].

[1] Throughout this paper we have used SHAP for the SHAPLEY approach

As highlighted in [10–16], recent efforts have aligned towards comprehensible and transparent model outputs. Fundamentally, the issue revolves around balancing the need for precise predictions with the ability to gain insight into the decision-making process, particularly with intricate models. A key tool used for the most interpretive model is the SHAP approach, which simply attributes the values of the predictors in a joint effort.

Despite the enhancements made in the SHAP-based approach as a feature selection technique, a substantial challenge exists in the selection of the best feature subset for detecting DDoS attacks.

As the model complexity changes through conditions like adding new feature, the stability of these attributions are critical , or it can shift in the assigned importance of features for the same problem. This is not just a technical issue; but also directly affects real-world decisions. Therefore, the complexity of the model impacts the SHAP-based feature selection, which is crucial for ensuring reliability, transparency, and informed decision-making systems.

In this approach, the Shap was incorporated to find the top features, that contributed maximumally to the classification. For model training, the CICDOS2017 dataset was used to train multiple machine learning models such as Logistic Regression, Random Forest, Decision Tree, Naïve Bayes, and neural network. In these models, Shap identifies the most significant features. This research proposes that the top features are retained across different models, but with slight changes in ranking.

To determine why the top feature rankings were inconsistent with respect to the machine learning model, a custom function was introduced. The function constitutes linear and non-linear terms, which can be scaled by factor α. The value of α ranged from 0 to 1. When α is zero, the model is completely non-linear, otherwise it is linear. The custom function is shown in (1)

$$\begin{aligned} f\{\nu(x_1, x_2)\} &= \alpha\{\nu(b_0 + b_1x_1 + b_2x_2)\}+ \\ &(1-\alpha)\{\nu(b_3x_1^2 + b_4x_1x_2 + b_5x_2^2+ \\ &b_6x_1^3 + b_7x_1^2x_2 + b_8x_1x_2^2 + b_9x_2^3)\} \end{aligned} \tag{1}$$

The highest degree of non-linearity was restricted to three, to reduce computational cost. By incorporating the Shap feature selection over the custom function, many variations in feature ranking were identified. There are multiple reasons for these variations in the top feature ranking, which were selected by Shap. The major reasons identified are model complexity and optimization properties.

The model optimization properties can be either convex or submodular. According to previous findings, if convex, then the Shap response will be consistent and aligned with the core, whereas the submodularity shows inconsistencies in feature ranking. Moreover, when the model has multiple local minima, the Shap variance is greater. In our proposed model, the nature of the model was identified and the Shap response was observwed with respect to the change in the model optimization properties. With the selected feature set, the model properties were identified, along with the Shap response.

The convexity of the model can be verified with the equation (2)

- **Convexity** –

$$f(S \cup \{i\}) - f(S) \leq f(T \cup \{i\}) - f(T) \tag{2}$$

- **w**here $S \subset T$ and $i \notin T$

The above equation states that, when the feature has been added to the large coalition, there is an increase in the feature contribution to the model prediction, On contrast, there is no visible change in the subset contribution.

Similarly, the submodularity of the model can be verified using equation (3)

- **Sub-Modularity** –

$$f(S \cup \{i\}) - f(S) \geq f(T \cup \{i\}) - f(T) \tag{3}$$

- **w**here $S \subset T$ and $i \notin T$

As shown in equation (3), when the new feature has been added to the large coalition, it does not show any change in the feature contribution to the model prediction. In contrast, the smaller coalition shows an appreciable change in the feature contribution towards model prediction.

Our contributions are summarized as follows 1. The model was integrated with the SHAP feature selection technique, and a comparative analysis was performed for the evaluation metrics. 2. A custom model was introduced to evaluate the SHAP performance of the linear and non-linear models. 3. Additionally, the SHAP approach was validated with changes in custom-model optimization properties such as convexity and submodularity.

The remainder parts of this article is organized as follows: Section II performs an in-depth exploration of pertinent literature. Our research strategy is described in the flow diagram in Section III. In section IV, we discuss the results obtained for different scenarios, and their comparison provides a better understanding of the system. Finally, Section V provides insights into a wide spectrum of exploration for future research.

2 Related Work

In this section, we provide a succinct summary of relevant research findings on the detection of DDoS attacks using different machine learning models followed by a deeper analysis of the SHAP uncertainties due to model properties.

2.1 DDoS attack detection using machine learning models

DDoS attacks are a more prevalent threat to various business networks, as they disrupts legitimate users from accessing the network. As highlighted in [7] [8], the approach focuses on monitoring the traffic from the botnets, which initiate attacks in normal traffic. Machine learning algorithms are utilized to perform insightful analyses and make decisions to identify attacks in their early stages.

In a recent study [1],researchers discussed an in-depth review of various DDoS detection approaches that include both mathematical and machine learning perspectives. This paper also reviews an array of machine learning algorithms, such as Random Forest and Convolutional Neural Networks.

In general, machine learning models include feature selection techniques to reduce the computational cost. These techniques assist in identifying the most important features for predicting a target. In a comprehensive review, researchers [17–20] explored the importance of selecting features in an explainable AI methodology based on the SHAP values. Unlike conventional approaches, such as filters, wrappers, and embedded, the features are selected based on their contribution to model prediction. This study research work focused on refining feature selection by integrating SHAP with other traditional feature selection approaches.

In reference [4, 21, 22], the model includes Explainable Artificial Intelligence (XAI) and SHAP to identify the key attributes for the classification. To combat DDoS attacks, experts have used machine learning and deep learning approaches. The framework includes the selection of top features using the SHAP technique and permutation feature importance. Additionally, SHAP provides detailed information on the contribution of the selected features to the model prediction.

2.2 Deeper Analysis of the SHAP Uncertainty

In [23], the unified SHAP approach, which is an explainable AI techniques, provides the level of feature contribution to the model. However, when the model properties changes, the SHAP approach exhibits some deviation in performance. The model property can be either convex or submodular, which plays a major role in many real-time applications. In [24], the algorithms are identified to test properties such as convexity and submodularity. Furthermore, in the literature [25], it was proved that submodular functions are relevant to machine learning models.

It was also found that the model properties strongly influence the SHAP performance. Moreover, in literature [26], if the game property is convex or totally convex, then the SHAP performance has been proven to be at the core.

Additionally, in an article [27], the concept of average convexity was proven for the weighted SHAP value, which aligns with the core. The notion of the SHAP value has been extended to the weighted SHAP, where features with a zero weight would also share their contribution.

However, there are a few other model properties, such as supermodular and superadditive properties. The model is supermodular, when the marginal contribution of adding a set of features is greater than that of performing individually. The superadditive property completely aligns with the submodular property. In [28], mathematical reasoning is provided to show that the supermodular and submodular are nothing but convex.

Similar to the above-mentioned few model optimization properties, convexity, submodularity, and superadditivity were identified as strongly influencing the SHAP performance on any machine learning model. Model complexity also plays a major role in determining feature pay-offs. In [29], the SHAP uncertainty was analyzed using a few machine learning models. Along with the performance metrics, the SHAP residuals

and SHAP variance were also observed for the mental model and the real machine learning model. In addition, the quantitative measure of uncertainity in the SHAP feature value was identified because of the randomness in the coalition, as studied in [30].

2.3 Validation of SHAP Performance

In conjunction with this, there is an interesting study on the application of SHAP to enhance the interpretability and explainability of model performance with a specific set of features. In paper [29], the paper introduced the Shapley Value Error Apportioning (SVEA), where the training error has been distributed to the feature which contributes less to the model prediction. The training error is determined using the hinge loss function. The feature that contributes more, will be given less error when compared with the feature that contributes very less to the model prediction. This study provides better interpretability and explainability of model performance for the set of features used for model prediction.

In [31], the alpha procedure has been introduced to check the feature contribution sequentially and it proves that SHAP's fairness rule works for machine learning algorithms. Moreover, SHAP approximation strategies when the dataset is large and tabular. In [32], the replacement and estimation strategies are clearly explained using the Kernel SHAP, TreeSHAP, and DeepSHAP methods. However, when the features are highly correlated SHAP may misrepresent true causality. In [30], the proposed work is based on a combined approach to preserve the feature dependencies and reduce the computational cost. Additionally, [33] also works on the quantitative measure of the SHAP uncertainity.

In [34, 35], the SHAP approach was used in their proposed framework, which makes the model more interpretable and explainable. A random forest classifier was used for target classification and enhanced the model transparency using the SHAP approach. Likewise, in [17], the SHAP approach was used as the feature selection technique, and it was integrated with LASSO to enhance performance.

In real-time applications, the researchers in papers [36–38] have used the SHAP approach for assessing the stability of the machine learning models and also to predict the target with the significant features.

In addition to the above research exploration, [34] introduces an open-source Python framework specifically to analyze the SHAP values for the joint contributions of different sets of features. The Shapiq framework was introduced to comprehend the contribution of the feature set together, unlike the traditional SHAP, which provides the SHAP value for the individual feature performance.

After an extensive literature survey, the research gap has been identified as to why SHAP uncertainties occur for the change in the model optimization properties, which motivated us to frame this proposed methodology.

As shown in Table 1, the literature has focused on convexity and submodularity using mathematical proofs. However, there are no findings on combinatorial optimization behaviour aligned with the machine learning model. In contrast, our proposed

Table 1 COMPARITIVE ANALYSIS OF EXISTING METHODS AND THE NOVEL PROPOSED FRAMEWORK

S. No.	Author(s), Year	Problem Addressed	Model	Dataset	Key Findings	Gaps Identified
1	SD. A. Ayoshin and T. Tanaka, 1999 [26]	The Shapley Value in Totally Convex Multichoice Games	No machine learning model	NA	Mathematical proof on convexity and submodularity	Limited to mathematical model; no machine learning model analysis
2	Michal Parnas and Dana Ron, 2002 [24]	On Testing Convexity and Submodularity	No machine learning model	NA	No SHAP variance analysis; focuses on convexity/submodularity testing	Convexity and submodularity mathematical proof; no machine learning model analysis
3	Y. Narahari, 2012 [28]	Game Theory	No machine learning model	No Dataset	Mathematical and theoretical proofs on convexity and submodularity	Convexity and submodularity mathematical proof; no machine learning model analysis
4	Francis Bach, 2013 [25]	Learning with Submodular Functions: A Convex Optimization Perspective	No machine learning model	NA	No SHAP variance analysis; focuses on submodular functions	Limited to optimization of submodular functions; Lovász extension provides regularization for supervised and unsupervised learning
5	Sandhya Tripathi, N. Hemachandra, Prashant Trivedi, 2021 [39]	Interpretable Feature Subset Selection: A Shapley Value Based Approach	SVEA Model	Thyroid, Pima Diabetes, Magic, Heart, IJCNN	Shapley value based error apportioning and feature identification; no combinatorial optimization analysis	Limited to interpretability and explainability
6	Ruiming Li, 2021 [29]	Quantifying Uncertainty in the Shapley-Value-Based Explanations for Machine Learning Models	Synthetic Model, XGBoost	NHANS, Synthetic	SHAP variance and SHAP residuals on custom equation	Limited to quantifying uncertainty in SHAP values; no convexity and submodularity check
7	Alexandre Skoda, Xavier Venel, 2023 [27]	Weighted Average-Convexity and Shapley Values	No machine learning model	No Dataset	Limited to weighted Shapley value and weighted convexity	Focus is on convexity and average convexity; no submodularity check
8	Antonella S. Antonini et al., 2024 [34]	Machine Learning Model Interpretability Using SHAP Values: Application to Igneous Rock Classification Task	Random Forest Classifier	90 surface samples, La Jovita-Las Aguilas mafic ultramafic belt, Argentina	Better classification using Random Forest with SHAP for model interpretation	Focus is on model interpretability using SHAP; no combinatorial optimization check
9	Maximilian Muschalik et al., 2024 [40]	shapiq: Shapley Interactions for Machine Learning	Decision Trees, Random Forest, KNN, Linear and Logistic Regression	Bike Sharing, California Housing, IMDB, Image Classification, Adult Census	Efficient computation of Shapley interactions for machine learning models; no SHAP variance or combinatorial optimization analysis	Focus is on Shapley value interactions
10	Adam Brandenburger et al., 2025 [31]	Reevaluating the Shapley Value: A New Justification and Extension	No Model	NA	Alpha procedure, uniqueness of Shapley value, weighting system	Limited to mathematical proofs; no combinatorial optimization check
11	Suchit Gupte and John Paparrizos, 2025 [32]	Shapley Value Approximation Techniques	Tree-Based and Neural Network Models	200 real and synthetic datasets	Focused on approximation techniques; model-specific approaches outperform model-agnostic methods	No combinatorial optimization check
12	Proposed Work	The SHAP deviations because of changes in model properties	Custom Model	CICIDS2017	Focused on finding SHAP variance due to change in model properties	Custom model used to check SHAP performance for change in model properties

work is a deeper analysis of the SHAP performance over changes in the model complexity and different feature sets. In additionally, the model was checked for its convexity and submodularity for different sets of features.

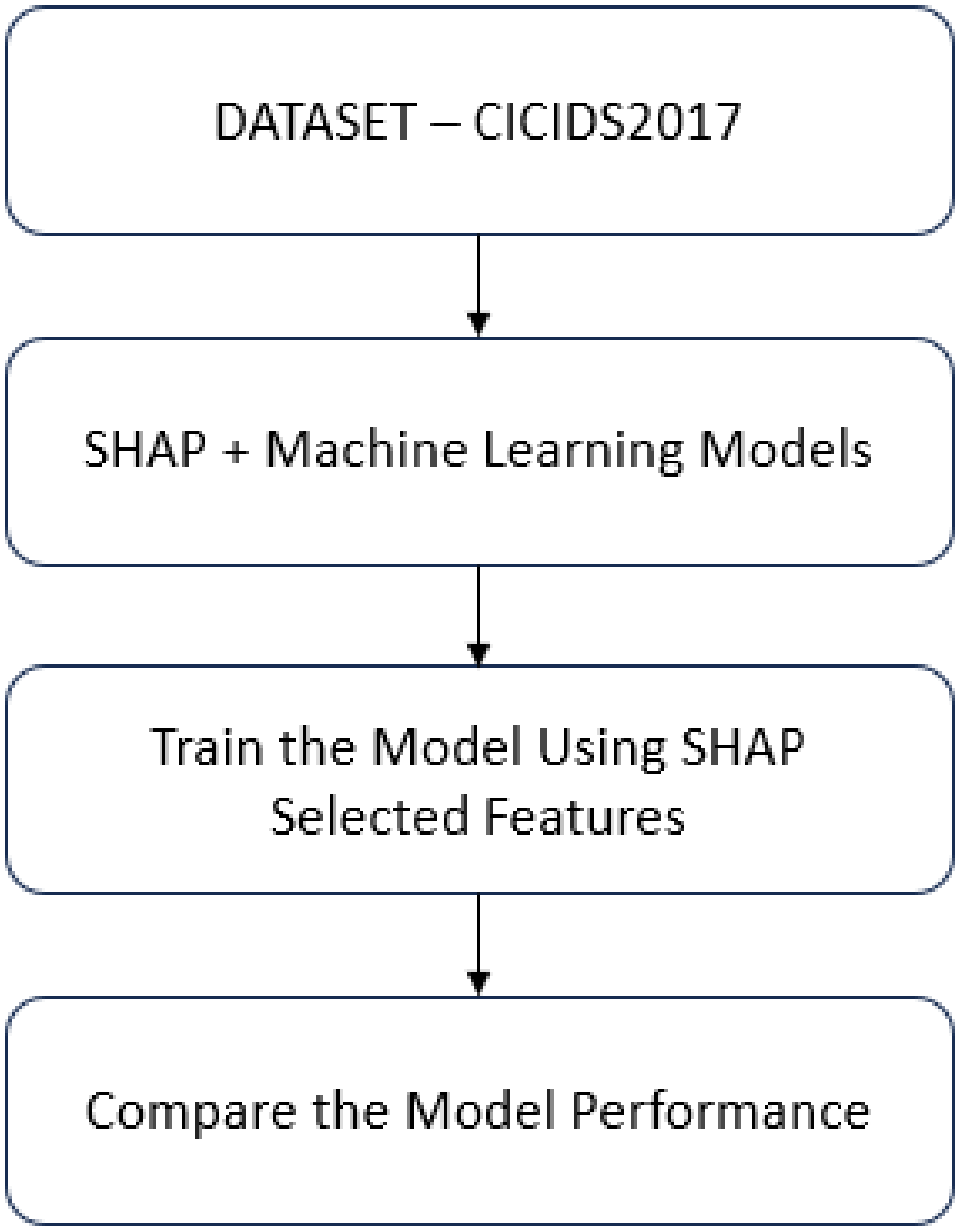


Fig. 1 PROPOSED FRAMEWORK.

3 Proposed Methodology

In this section, we discuss the proposed framework for detecting DDoS attacks using the SHAP feature selection technique, as shown in Figure 1 . Moreover, we present an analysis of why SHAP shows few inconsistencies in performance.

3.1 Data analysis

The CICIDS2017 dataset is from the Kaggle repository was used for the analysis of DDoS attacks . The dataset was created by the researchers using the CICFlow Meter tool. The tool is preferably used to capture network traffic and the features are used later to train the machine learning model. It can also be used for various other tasks, such as network traffic analysis, anomaly detection, and consistent monitoring of the network to maintain a high level of security. It provides a clear insight into the deviations in network traffic, and it is an alarming factor for network administrators to protect the system from anomalies.

Intuitively, the dataset can be analyzed, and the deviations of every feature can be found as per the traffic record to determine whether the attack has been initiated. There are multiple features such as destination port, initial window forward bytes, forward packet length max, backward packet length max, inter-arrival time, and so on.

For example, a few features such as backward packet length max, forward packet length max, initial window size, subflow forward bytes, and destination port are analyzed directly from the dataset. Backward packet length max shows the significant deviation as 720.7 for benign and 3963.24 for malicious. Similarly, for subflow forward bytes the standard deviation is 4738.38 for benign and 11.83 for malicious. Likewise, for the destination port, the deviation is 25633.94 for benign, and whereas 0 for malicious. From the examples above, it is clear that the difference between normal and malicious traffic is significantly higher. This can be cross-validated using other correlation tests.

Statistically, a few correlation tests can provide a proper understanding of the relationship between the various features under consideration. A few such correlation tests are ANOVA, t-test, Mann-Whitney U test, and Kruskal-Wallis test. The tests are categorized based on two approaches, parametric and non-parametric approaches. If the data did not follow any distribution, then Mann-Whitney and Kruskal-Wallis tests were preferred to study the statistical similarities between the groups. In both cases, the p-value was less than 0.005, and the null hypothesis. The inference of rejecting the null hypothesis is that there are significant differences between the groups.

Statistical tests were performed between benign and malicious features to identify abnormalities in network traffic records. From the correlation tests, it is much easier to comprehend the deviation from normal flow. Thus, these are the few analyses that have been conducted both intuitively and statistically.

3.1.1 ANDREWS CURVES

Exploratory Data Analysis is one of the most important steps in analyzing the data structure, revealing patterns, outliers, and other statistical measures. Andrews curves are one of the best ways to represent high dimensional data through the conversion of each data point into a sine and cosine function series spanning between $-\pi$ and π.

This technique visualizes distinct classes within a dataset by plotting separate curves for every individual data point.

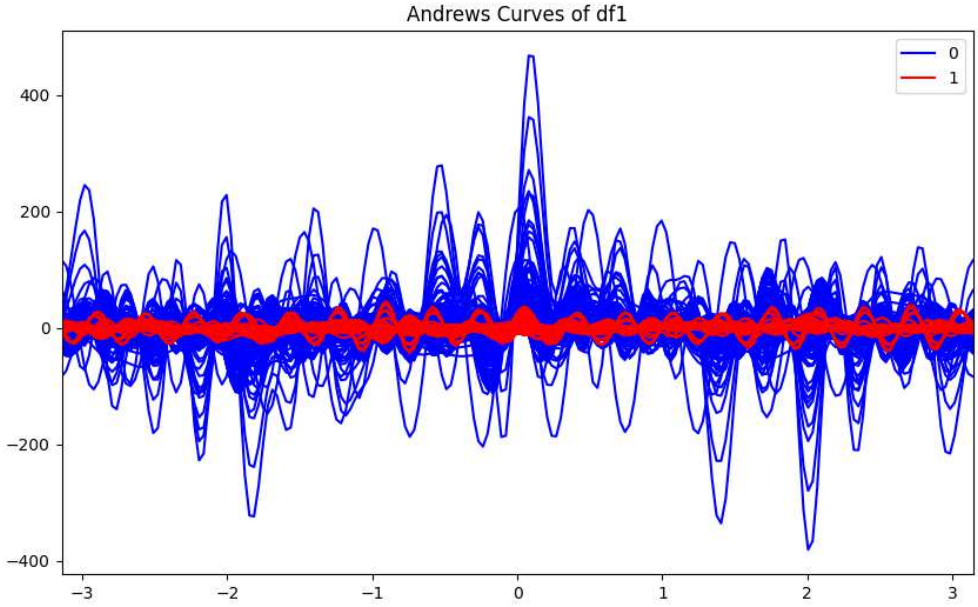


Fig. 2 ANDREWS CURVES - CICIDS2017.

From Figure 2 we infer that the binary classification data have overlapping points, which reveals non-linearity in the dataset. The presence of overlapping points is a clear indication of the inseparable boundaries between classes in the data.

3.2 System Architecture

As shown in Figure 1, the proposed framework includes the machine learning model integrated with the SHAP feature selection technique for DDoS attack detection. This architecture shows how SHAP performance varies with changes in the model optimization properties.

First, the SHAP selected features were used in conventional machine learning models such as random forest classifier, neural network, and logistic regression for prediction. Algorithm 1 provides the detailed steps of the top feature selection and the SHAP variance for the different machine learning models.

In addition, the custom model was analyzed with a different set of features to check the SHAP performance on multiple levels of model complexities. As shown in Equation(1), the custom model evaluation metrics and the SHAP variance were analyzed.

A few sets of features were identified using the SHAP feature selection technique for DDoS attack detection. However, when the model changes, the SHAP approach

Algorithm 1 Feature Selection and SHAP Variance

Require: Dataset X, labels y, model M, number of iterations N

Let initial feature set $X^{(0)} = \{1, 2, \ldots, N\}$

$N = totalfeatures$

For each feature X_i, set $Value[i] \leftarrow 0$, $ShapVar[i] \leftarrow 0$

for $t = 1$ to N **do**

Randomly split (X, y) into (X_{train}, y_{train}) and (X_{test}, y_{test})

Train model on current feature set $X^{(t-1)}$:

$$M^{(t)} \leftarrow M(X_{train}[X^{(t-1)}], y_{train})$$

Compute SHAP values

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [v(S \cup \{i\}) - v(S)]$$

$\forall X_J \in X_{test}$ and $i \in X^{(t-1)}$ for each feature $X_i \in X^{(t-1)}$

compute the mean SHAP value in iteration t as follows:

$$\overline{\phi_i}^{(t)} = \frac{1}{|X_{test}|} \sum_{j=1}^{|X_{test}|} \phi_i^{(t)}(X_j)$$

Compute SHAP variance for feature i in iteration t:

$$\text{Shap Var}^{(t)}(\phi_i) = \frac{1}{|X_{test}|} \sum_{j=1}^{|X_{test}|} \left(\phi_i^{(t)}(x_j) - \overline{\phi_i}^{(t)}\right)^2$$

Accumulate SHAP importance across iterations:

$$ValScore[i] \leftarrow ValScore[i] + \left|\overline{\phi_i}^{(t)}\right|$$

Accumulate SHAP variance across iterations:

$ShapVar[i] \leftarrow ShapVar[i] + ShapVar^{(t)}[\phi_i]$

end for

Update feature set $X_i^{(t)}$ by keeping features with the highest $Val_Score[i]$ and lowest $Shap_Var[i]$ to

compute the average SHAP importance and SHAP variance: $ValScore[i] \leftarrow \frac{ValScore[i]}{N}$, $ShapVar[i] \leftarrow \frac{ShapVar[i]}{N}$

Rank all features X_i using:

$$\text{High } ValScore[i] \text{ and Low } ShapVar[i]$$

Output: Top features based on combined High SHAP importance and Low SHAP variance

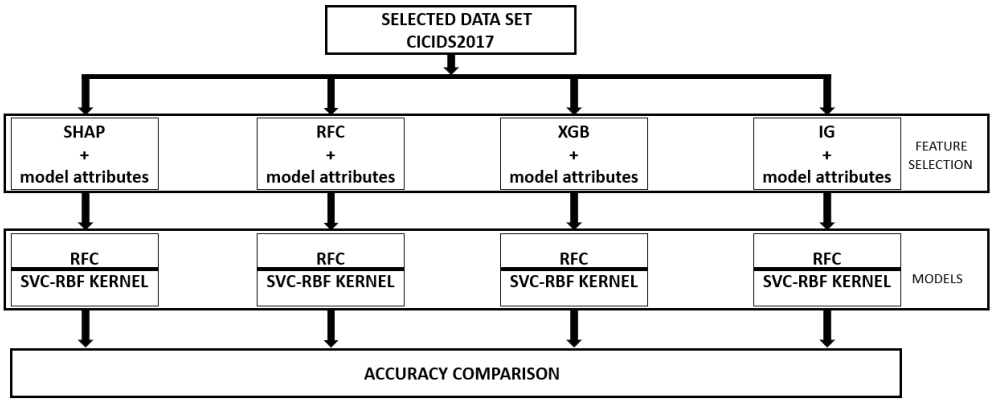


Fig. 3 SHAP vs. other feature selection techniques.

exhibits some deviations in performance. Further analysis, revealed that the SHAP performance changed with the changes in the model properties. The following section provides detailed explanations of the model properties for the different sets of features, used for training.

3.3 Model Property Analysis

Following the SHAP integration with the machine learning model for prediction, the model properties were identified, and the SHAP variance was analyzed. In general, the model properties are convex, submodular, and superadditive. According to [26], when the model is convex, the SHAP is in the core. Algorithm 2 provides the steps for identifying whether the model is convex or submodular.

4 Results and Discussions

4.1 SHAP Vs Other Feature Selection Techniques

The datasets incorporated in our research analysis are CICIDS2017 / Friday Working Hours Morning (BoT Attack), CICIDS2017 / Friday Working Hours Afternoon (DDoS attack), and CICEV2023 (DDoS attack on electric vehicles).

Table 2 Performance metrics comparison on various datasets

Dataset	Accuracy	F1	Precision	Recall
CICIDS2017 (DDoS)	0.98	0.99	0.98	0.98
CICIDS2017 (Bot)	0.98	0.99	0.98	0.98
CICEV2023	0.98	0.99	0.98	0.98

Algorithm 2 Model Complexity Vs Shap Variance

Require: Dataset X, feature set $F = \{1, \ldots, N\}$, convexity tests C_N, submodularity tests S_N, number of SHAP runs N

Define the model

$$f\{\nu(x_1, x_2)\} = \alpha\text{linear part} + (1-\alpha)\text{non-linear part}$$

Set flags: is_convex $\leftarrow$ true, is_submodular $\leftarrow$ true

Convexity Check
Sample two features: x_1, x_2
for $t = 1$ to C_N **do**
    Sample feature set x_1, x_2 from X
    $\alpha \in [0, 1]$
    Compute

$$f\{\nu(x_1, x_2)\} = \alpha\{\nu(b_0+b_1x_1+b_2x_2)\}+(1-\alpha)\{\nu(b_3x_1^2+b_4x_1x_2+b_5x_2^2+b_6x_1^3+b_7x_1^2x_2+b_8x_1x_2^2+b_9x_2^3)\}$$

    Check convexity inequality: $f(S\cup\{i\})-f(S) \leq f(T\cup\{i\})-f(T)$ where $S \subset T$ and $i \notin T$
    if inequality is not satisfied **then**
        is_convex $\leftarrow$ false
    end if
end for

Submodularity Check
for $t = 1$ to S_N **do Sample two features:** x_1, x_2 **Compute the Model**
    Submodularity Check
    Check Submodularity In-Equality $f(S\cup\{i\})-f(S) \geq f(T\cup\{i\})-f(T)$ where $S \subset T$ and $i \notin T$
    if inequality is violated **then**
        is_submodular $\leftarrow$ false
    end if
end for

Compute SHAP Variance
Initialize SHAP storage: $\phi_i^{(N)} \leftarrow 0$ for all $i \in N$, $t = 1, \ldots, N$
for $k = 1$ to N **do**
    Feature Subset $X_i \subset X$
    Evaluate the model f on X_i
    Compute SHAP values $\phi_i^{(N)}$ for all $i \in N$
end for
for each feature $i \in N$ **do**
    Compute mean SHAP across runs:

$$\overline{\phi_i} = \frac{1}{N}\sum_{N=1}^{N}\phi_i^{(N)}$$

    Compute SHAP variance:

$$\text{Shap Var}(\phi_i) = \frac{1}{N}\sum_{N=1}^{N}\left(\phi_i^{(N)} - \overline{\phi_i}\right)^2$$

end for
Output: is_convex, is_submodular, and Shap Var(ϕ_i) for all $i \in N$

The table 2 compares the performance metrics of the proposed methodology while using on various datasets.

The heuristic approach includes different machine learning models along with the SHAP approach for significant feature selection, as shown in Figure 3. Other feature selection techniques such as RFC (Random Forest Classifier), XGB (Extreme Gradient Boosting), and IG (Information Gain) were also used for selecting features and compared with the SHAP-selected feature performance.

The observations of our research are articulated in four distinct tables.

Table 3 ACCURACY COMPARISON

ML Model	2 Features	5 Features	10 Features	20 Features
RFC	99.96	99.94	99.88	99.96
XGB	99.92	100	100	99.99
NN	99.38	99.84	99.8	99.75
LR	98	98.5	98.7	98.8
Custom Model	98	98.5	99	99

Table 4 F1-SCORE COMPARISON

ML Model	2 Features	5 Features	10 Features	20 Features
RFC	99.87	99.94	99.86	99.96
XGB	99.82	99.9	99.9	99.98
NN	99.38	99.84	99.78	99.75
LR	98	98.65	98.8	98.8
Custom Model	98	98.5	98.6	98.7

Table 3 and 4 show a comprehensive comparison of all performance metrics such as the accuracy and F1-Score for different sets of SHAP-based selected features on test data. The performance metrics include the accuracy and F1-Score, which provides a complete spectrum assessment of the model's performance. The formulas for the accuracy and F1-score are shown in Equation 4 and Equation 5 respectively.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{4}$$

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

In Equation 5, precision is the ratio of correctly classified DDoS attacks to the total predicted DDoS attacks, and recall is the ratio of correctly predicted DDoS to the total actual DDoS attacks.

For instance, consider the RFC model where the metrics are observed for different feature subsets. Figure 4 provides a better visualization of the scores for different iterations.

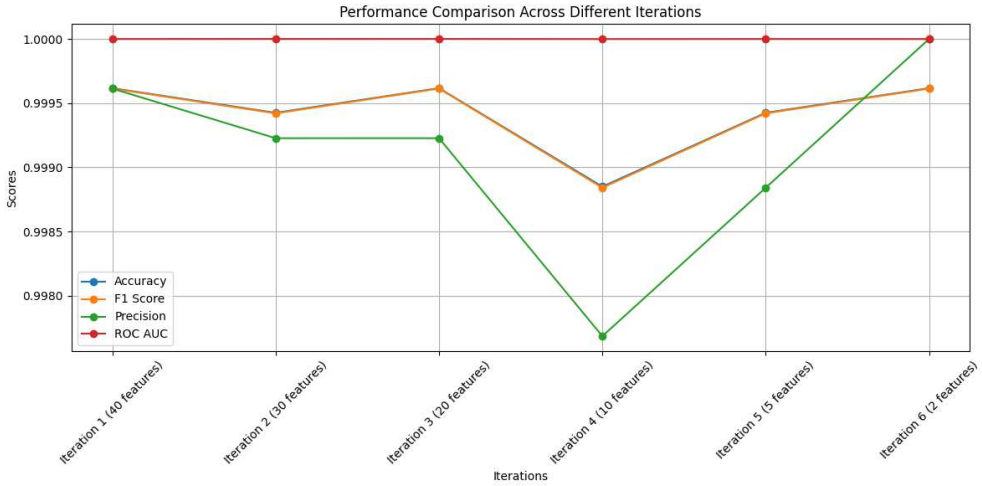


Fig. 4 RFC PERFORMANCE METRIC ON VARIOUS FEATURE SUBSETS.

Table 5 SENSITIVITY COMPARISON

ML Model	2 Features	5 Features	10 Features	20 Features
RFC	99.85	99.85	99.86	99.8
XGB	99.96	99.72	100	100
NN	100	100	100	100
LR	100	100	100	100
Custom Model	99	99.5	99.7	99

Table 6 SPECIFICITY COMPARISON

ML Model	2 Features	5 Features	10 Features	20 Features
RFC	90	86.3	87	86
XGB	96	99.88	99.96	100
NN	99	100	100	100
LR	94	97	98	100
Custom Model	97	96	98	99

Table 5 and 6 compare the true positive rate, which is otherwise known as the sensitivity, and the true negative rate, which is also known as the specificity. Sensitivity is the measure of the model correctly identifying the malicious instances. Specificity assesses the critical discrimination of benign instances. These measures are important for validating the efficacy of the model in discriminating benign instances from malicious instances.

In further analysis, the two most significant features are selected by other feature selection techniques, such as the Random Forest Classifier (RFC), XGBoost (XGB), and Information Gain (IG), and are validated by training two different models. The

Table 7 OTHER FEATURE SELECTION TECHNIQUES

Other Techniques	Model 1 - RFC (Accuracy)	Model 2 - SVC-RBF Kernel (Accuracy)
RFC	90	86
XGB	98	78
IG	98	85

Table 8 SHAP FEATURE SELECTION TECHNIQUES

SHAP Feature Selection	Model 1 - RFC (Accuracy)	Model 2 - SVC-RBF Kernel (Accuracy)
RFC	98	96
SVC-RBF	99	93

models selected for evaluating the selected features were the RFC and support vector classifier – RBF Kernel. Later, the same comparison was made for the two most significant features, which were based on the SHAP-values for the respective machine learning model.

Tables 7 and 8 show the accuracy scores for different models, which incorporates the two most significant features selected by two distinct methodologies: conventional feature selection techniques and SHAP-based feature selection. Table 7 shows inconsistencies in the accuracy score for different models, whereas Table 8, shows that irrespective of the models the SHAP selected features are highly efficient in prediction. The accuracy score for the SHAP-based approach was substantially high (¿ 90 percent) for the different models. However, the features selected by other techniques are model-specific, and thus there is a noticeable decrease in accuracy.

4.2 Comparison With Existing Literature

Table 9 presents a comparative analysis of the existing literature. To begin with [2], the researchers used the CICDDoS2019 dataset and SHAP recursive feature elimination with cross-validation as feature selection, which is integrated with classifiers such as Decision Tree, XGB, LGB, and RF. The accuracy score was 99.98% four selected features, whereas the accuracy decreased to 92.2% for two features. Subsequently, [4] incorporated XGB-based feature selection on the CICDDoS2019 dataset and SHAP for the interpretability of the model, which attained 99% accuracy for 20 features. Furthermore, [6] utilized one hot encoding and robust scalar for feature selection over the UNSWNB15 dataset with 94.14% accuracy on training data. Additionally, [8] incorporated J48 as a feature selection technique and acheived 98.6% accuracy. Finally, [3] incorporated the ExtraTree classifier as feature selection technique over the BOT IOT dataset and attained an accuracy score of 100% for the 15 selected features. In our proposed model, the accuracy score is consistent at approximately 98% with two

Table 9 Proposed methodology vs. existing literature findings for DDoS attack detection

Ref	**Feature Selection**	**No. of Features**	**Classifier**	**Accuracy**	**Consistency Check**
[2]	SHAP RFECV	4	KNORA-E/U	99.9	X
[4]	XGB	20	SHAP	99	X
[6]	One-hot Encoding	30	RF, KNN, SVM, DT	91.88	X
[8]	J48	27	WEKA	98.6	X
[3]	Extra-Tree	15	RF, NB	100	X
Proposed	SHAP	2	RF, XGB, LR	99	✓

significant features, by using SHAP as feature selection technique to retrain the model with dynamically changing feature subsets based on the model performance.

4.3 Custom Model versus SHAP uncertainities

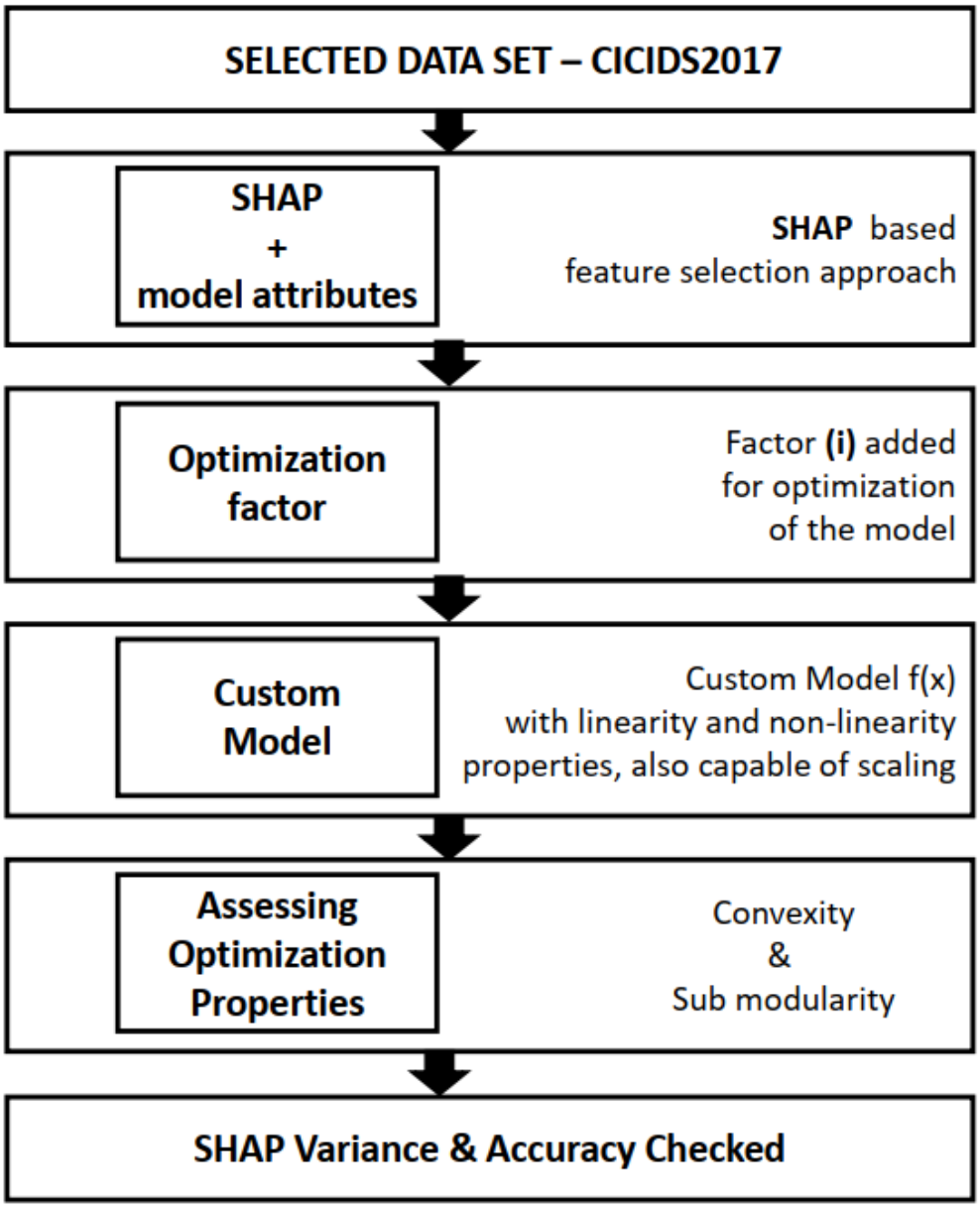


Fig. 5 Custom Model Vs SHAP Uncertainities.

In this study, the custom model was trained with multiple subsets of features, and checked whether the model was convex or submodular, as shown in Figure 5. It shows submodularity for the selected set of features and the model's Shap variance was high for non-linear with submodular properties which ranged between 2000 and 17,000. Furthermore, when the model was trained using a different set of features, it was identified as convex. The Shap variance was found to be between 10 and 30 for non-linear with convex properties. The accuracy was maintained at a high for the same. Thus, the experimental results also prove the aforementioned Shapley variance,

Additionally, when the number of features was increased to 10 and the additional feature 'i', which was considered as one of the most important features, was added to the coalition, the model showed submodularity for α range 0.7 to 1 and convexity for the range 0 to 0.6. Consequently, the Shapley variance was high for both the linear and non-linear parts of the model range was between 2000 - 8000. In addition, the accuracy was consistently high for α range of 0 to 0.7. Thus, the impactness of a feature plays major role in model performance for different game properties.

The inference from the heuristic approach is that SHAP is influenced not only by the model complexity but also by the set of features used for the training. Only if all of these features in the set, contribute to the model prediction, irrespective of the model complexity, the SHAP variance will be high because all these selected features contribute significantly to the model prediction for different optimization properties.

Consider:

Large Dataset: T

$$T = \{x_1, x_2, x_3\}$$

Small Dataset: S

$$S = \{x_1, x_2\}$$

i - Additional Feature

Custom Function for the smaller set: Let degree = 2

$$f\{\nu(x_1, x_2)\} = \nu\{\alpha(b_0 + b_1x_1 + b_2x_2)\} + \nu\{(1-\alpha)(b_3x_1^2 + b_4x_1x_2 + b_5x_2^2)\} \tag{6}$$

Custom Function for the larger set:

Let degree = 2

$$f\{\nu(x_1, x_2, x_3)\} = \nu\{\alpha(b_0 + b_1x_1 + b_2x_2 + b_3x_3)\} + \nu\{(1-\alpha)(b_4x_1^2 + b_5x_2^2 + b_6x_3^2 + b_7x_1x_2 + b_8x_2x_3 + b_9x_3x_1)\} \tag{7}$$

Add feature i to S and T, where $i \notin \{S, T\}$.

$$f\{\nu(T, S)\} = \begin{cases} \text{Convex}, & \text{if } (S \cup i) < (T \cup i), \\ \text{Submodular}, & \text{otherwise.} \end{cases}$$

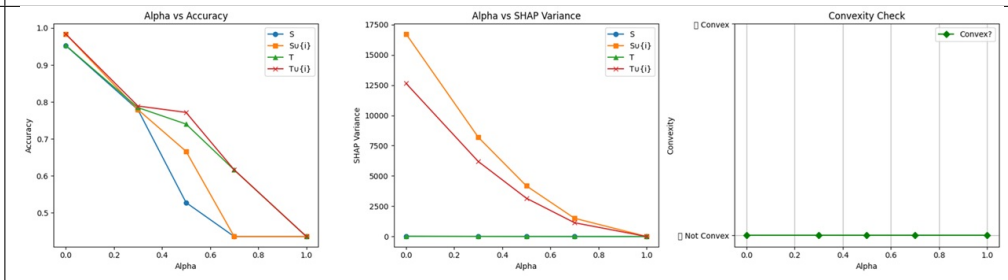
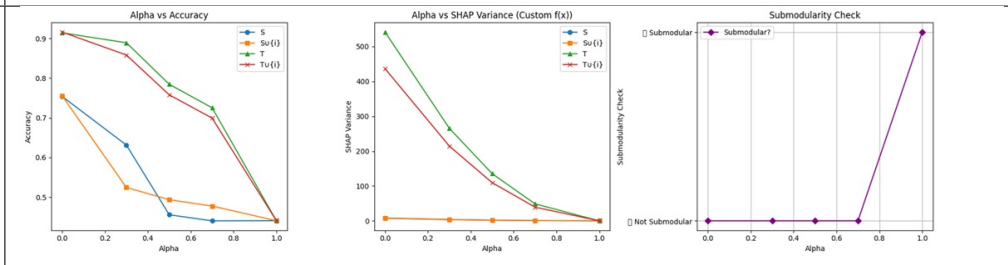
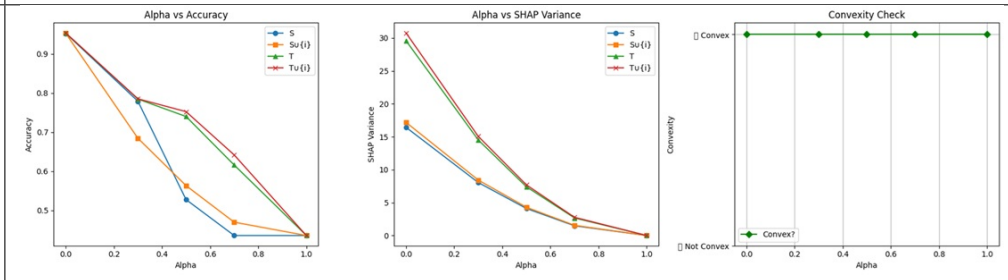
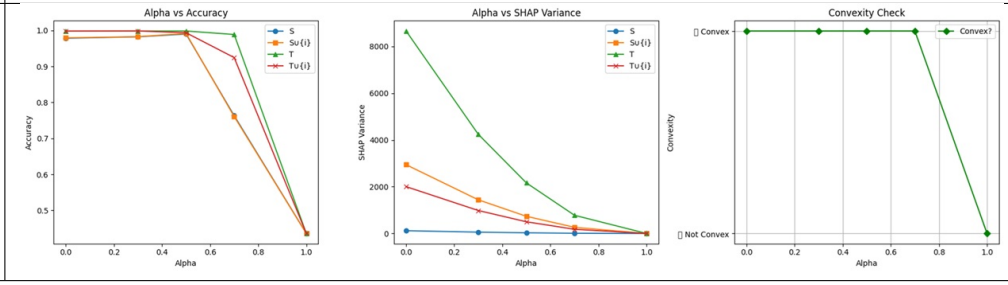
$$f(\{T, S\} \cup i) = \begin{cases} \text{Property Preserved}, & \text{'i' less significant}, \\ \text{Property Changes}, & \text{otherwise.} \end{cases} \tag{8}$$

As shown in Table 10, the model optimization properties, accuracy, and Shap variance for different sets of features were analyzed. The Shap variance was high for the submodularity pattern. Additionally, for further comprehension, the sets 1 and 3 were given the same set of features for both smaller and larger coalitions with a change in additional features. Consequently, the model properties of the feature subsets were different. From this, it is more clear that the importance of the feature influences the model optimization properties and Shapley variance. The images of the results obtained for the multiple feature sets are displayed in Table 11.

Table 10 DIFFFERNT GAME PARTICIPANTS

Sets	**Actual Set**	**Sub Set**	**Additional Feature (i)**	**Pattern**	**Highest Coalition Accuracy**	**Shap Variance**
Set 1	{'Destination Port', 'Avg Bwd Segment Size', 'Packet Length Variance'}	{'Destination Port', 'Avg Bwd Segment Size'}	act_data_pkt _fwd	Submodular	$(S \cup \{i\}) = 98$	12k to 17k
Set 2	{'Destination Port', 'Idle Min', 'Packet Length Variance', 'Fwd Packets/s', 'Flow IAT Max', 'Fwd IAT Min'}	{'Destination Port', 'Idle Min'}	Idle Std	Submodular	$(T \cup \{i\}) = 95$	4k to 5k
Set 3	{'Destination Port', 'Avg Bwd Segment Size', 'Packet Length Variance'}	{'Destination Port', 'Avg Bwd Segment Size'}	Idle Std	Convex	$(T \cup \{i\})$, T, S, $(S \cup \{i\}) = 96$	16 to 30
‘ Set 4	{'Average Packet Size', 'Destination Port', 'Packet Length Variance', 'Min Packet Length', 'Bwd Packet Length Min', 'Avg Bwd Segment Size', 'act_data_pkt _fwd', 'min_seg_size _forward', 'Flow IAT Mean','Init _Win_bytes _forward'}	{'Average Packet Size', 'Destination Port', 'Packet Length Variance','Min Packet Length', 'Init_Win _bytes_forward'}	Active Max	Convex for non-linear and Submodular for linear	$(T \cup \{i\}) = 100$	2k to 8k

Table 11 Feature Sets with Corresponding SHAP Variance, Accuracy, and Optimization Properties

Feature Set	Image
Set 1	
Set 2	
Set 3	
Set 4	

5 Conclusion and Future Work

The research presented in this study introduces a novel concept for assessing the performance of SHAP in integration with different machine learning models. The efficiency of the SHAP approach was validated by determining the model optimization properties. Likewise, the model properties are checked for the custom model f(x) using different feature subsets. Furthermore, the SHAP variance was found to be lower for linear and convex properties. Additionally, it was identified that the model property continues to change with the changes in the feature set. Consequently, the heuristic approach proves that the SHAP performance is influenced by the model optimization properties.

It is evident that the attribution-driven pipeline exhibits significant inconsistencies in feature selection under varying model complexities. This indicates that a reasonable level of error tolerance must be anticipated for reliable and informed decision making in high-stake fields such as cybersecurity, medicine, and the financial sector etc.

Several future research directions have emerged from this study. One is to further explore the operation of the SHAP in a real-time environment and assess its consistency. Finally, another potential research avenue would be to extend the work with the SHAP over a deep neural network architecture and check how it responds to optimization complexities.

6 Author Contributions

Cynthia C conceived the study, developed the proposed custom model, performed the analysis, and wrote the main manuscript. Gopal Krishna Kamath, Debayani Ghosh, and Prashant K. Wali reviewed the manuscript. All authors read and approved the final manuscript.

7 Acknowledgments

The author would like to acknowledge the support and facilities provided by BITS Pilani, Hyderabad campus and Christ University, Kengeri Campus, for carrying out this work seamlessly.

8 Research Funding Statement

The authors declare that no funding was received for the research, authorship, and/or publication of this manuscript

References

[1] Singh, S., Gupta, M., Sharma, D.K.: Ddos attack detection with machine learning: A systematic mapping of literature. In: 2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT), pp. 939–945 (2023). IEEE

[2] Batchu, R.K., Seetha, H.: An integrated approach explaining the detection of distributed denial of service attacks. Computer Networks **216**, 109269 (2022)

[3] Sadhwani, S., Manibalan, B., Muthalagu, R., Pawar, P.: A lightweight model for ddos attack detection using machine learning techniques. Applied Sciences **13**(17), 9937 (2023)

[4] Wei, Y., Jang-Jaccard, J., Singh, A., Sabrina, F., Camtepe, S.: Classification and explanation of distributed denial-of-service (ddos) attack detection using machine learning and shapley additive explanation (shap) methods. arXiv preprint arXiv:2306.17190 (2023)

[5] Golduzian, A.: Predict and prevent ddos attacks using machine learning and statistical algorithms. arXiv preprint arXiv:2308.15674 (2023)

[6] Najar, A.A., Naik, S.M.: Applying supervised machine learning techniques to detect ddos attacks. In: 2022 2nd Asian Conference on Innovation in Technology (ASIANCON), pp. 1–7 (2022). IEEE

[7] Satyanarayana, D., Alasmi, A.S.: Detection and mitigation of ddos based attacks using machine learning algorithm. In: 2022 International Conference on Cyber Resilience (ICCR), pp. 1–5 (2022). IEEE

[8] Saini, P.S., Behal, S., Bhatia, S.: Detection of ddos attacks using machine learning algorithms. In: 2020 7th International Conference on Computing for Sustainable Global Development (INDIACom), pp. 16–21 (2020). IEEE

[9] Devi, R.S., Bharathi, R., Kumar, P.K.: Investigation on efficient machine learning algorithm for ddos attack detection. In: 2023 International Conference on Computer, Electrical & Communication Engineering (ICCECE), pp. 1–5 (2023). IEEE

[10] Chen, H., Lundberg, S.M., Lee, S.-I.: Explaining a series of models by propagating shapley values. Nature communications **13**(1), 4512 (2022)

[11] Guardiola, L.A., Meca, A.: The shapley value and the strength of weak players in big boss games. International Journal of Game Theory **55**(1), 18 (2026)

[12] Sundararajan, M., Najmi, A.: The many shapley values for model explanation. In: International Conference on Machine Learning, pp. 9269–9278 (2020). PMLR

[13] Chen, H., Covert, I.C., Lundberg, S.M., Lee, S.-I.: Algorithms to estimate shapley value feature attributions. Nature Machine Intelligence **5**(6), 590–601 (2023)

[14] Merrick, L., Taly, A.: The explanation game: Explaining machine learning models using shapley values. In: International Cross-domain Conference for Machine Learning and Knowledge Extraction, pp. 17–38 (2020). Springer

[15] Rodríguez-Pérez, R., Bajorath, J.: Interpretation of compound activity predictions from complex machine learning models using local approximations and shapley values. Journal of medicinal chemistry **63**(16), 8761–8777 (2019)

[16] Henninger, M., Strobl, C.: Interpreting machine learning predictions with lime and shapley values: theoretical insights, challenges, and meaningful interpretations. Behaviormetrika **52**(1), 45–75 (2025)

[17] Gramegna, A., Giudici, P.: Shapley feature selection. FinTech **1**(1), 72–80 (2022)

[18] Smith, M., Alvarez, F.: Identifying mortality factors from machine learning using shapley values–a case of covid19. Expert Systems with Applications **176**, 114832 (2021)

[19] Jia, R., Dao, D., Wang, B., Hubis, F.A., Hynes, N., Gürel, N.M., Li, B., Zhang, C., Song, D., Spanos, C.J.: Towards efficient data valuation based on the shapley value. In: The 22nd International Conference on Artificial Intelligence and Statistics, pp. 1167–1176 (2019). PMLR

[20] Feretzakis, G., Sakagianni, A., Anastasiou, A., Kapogianni, I., Bazakidou, E., Koufopoulos, P., Koumpouros, Y., Koufopoulou, C., Kaldis, V., Verykios, V.S.: Integrating shapley values into machine learning techniques for enhanced predictions of hospital admissions. Applied Sciences **14**(13), 5925 (2024)

[21] Louhichi, M., Nesmaoui, R., Mbarek, M., Lazaar, M.: Shapley values for explaining the black box nature of machine learning model clustering. Procedia computer science **220**, 806–811 (2023)

[22] Jamshidi, P., Nowaczyk, S., Rahat, M.: Analysis of characteristic functions on shapley values in machine learning. In: 2024 International Conference on Intelligent Environments (IE), pp. 70–77 (2024). IEEE

[23] Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. Advances in neural information processing systems **30** (2017)

[24] Parnas, M., Ron, D., Rubinfeld, R.: On testing convexity and submodularity. SIAM Journal on Computing **32**(5), 1158–1184 (2003)

[25] Bach, F.: Learning with submodular functions: A convex optimization perspective. arXiv preprint arXiv:1111.6453 (2011)

[26] Ayoshin, D.A., Tanaka, T.: The shapley value in totally convex multichoice games. Applied Mathematics Letters **13**(4), 95–98 (2000)

[27] Skoda, A., Venel, X.: Weighted average-convexity and shapley values. Games and Economic Behavior **140**, 88–98 (2023)

[28] Chalkiadakis, G., Elkind, E., Wooldridge, M.J.: Computational Aspects of

Cooperative Game Theory. Springer, ??? (2012)

[29] Li, R.: Quantifying uncertainty in shapley-value-based explanations for machine learning models. PhD thesis (2021)

[30] Kargin, V.: Uncertainty of the shapley value. International game theory review **7**(04), 517–529 (2005)

[31] Brandenburger, A., Nalebuff, B.: Reevaluating the shapley value: a new justification and extension. Available at adambrandenburger. com/aux/material/Shapley (2023)

[32] Gupte, S., Paparrizos, J.: Understanding the black box: A deep empirical dive into shapley value approximations for tabular data. Proceedings of the ACM on Management of Data **3**(3), 1–31 (2025)

[33] Rozemberczki, B., Watson, L., Bayer, P., Yang, H.-T., Kiss, O., Nilsson, S., Sarkar, R.: The shapley value in machine learning. In: The 31st International Joint Conference on Artificial Intelligence and the 25th European Conference on Artificial Intelligence, pp. 5572–5579 (2022). International Joint Conferences on Artificial Intelligence Organization

[34] Antonini, A.S., Tanzola, J., Asiain, L., Ferracutti, G.R., Castro, S.M., Bjerg, E.A., Ganuza, M.L.: Machine learning model interpretability using shap values: Application to igneous rock classification task. Applied Computing and Geosciences **23**, 100178 (2024)

[35] Pandl, K.D., Feiland, F., Thiebes, S., Sunyaev, A.: Trustworthy machine learning for health care: scalable data valuation with the shapley value. In: Proceedings of the Conference on Health, Inference, and Learning, pp. 47–57 (2021)

[36] Han, T., Chen, J., Wang, L., Cai, Y., Wang, C.: Interpretation of stability assessment machine learning models based on shapley value. In: 2019 IEEE 3rd Conference on Energy Internet and Energy System Integration (EI2), pp. 243–247 (2019). IEEE

[37] Li, M., Sun, H., Huang, Y., Chen, H.: Shapley value: from cooperative game to explainable artificial intelligence. Autonomous Intelligent Systems **4**(1), 2 (2024)

[38] Nohara, Y., Matsumoto, K., Soejima, H., Nakashima, N.: Explanation of machine learning models using shapley additive explanation and application for real data in hospital. Computer Methods and Programs in Biomedicine **214**, 106584 (2022)

[39] Tripathi, S., Hemachandra, N., Trivedi, P.: Interpretable feature subset selection: A shapley value based approach. In: 2020 IEEE International Conference on Big Data (Big Data), pp. 5463–5472 (2020). IEEE

[40] Muschalik, M., Baniecki, H., Fumagalli, F., Kolpaczki, P., Hammer, B., Hüllermeier, E.: shapiq: Shapley interactions for machine learning. Advances in Neural Information Processing Systems **37**, 130324–130357 (2024)