Improving Anomaly Detection in Network Traffic Using Choquet-Based Feature Engineering for Random Forest and XGBoost Models

Abreu Quevedo[1], Denner Ayres[1], Gabriel Teixeira[1], Graçaliz Dimuro[1], Giancarlo Lucca[2], and Bruno L. Dalmazo[1(✉)]

[1] Federal University of Rio Grande, Rio Grand, Brazil
{abreu_rg,dennerayres,gabsi,gracaliz,dalmazo}@furg.br
[2] University Católica of Pelotas, Pelotas, Brazil
giancarlo.lucca@ucpel.edu.br

Abstract. Network traffic is essential for modern communication, ensuring the proper functioning of daily activities. In today's connected world, cybercriminals attempt to harm and extort users. To address this issue, various models have been proposed, yet they still underperform. This study proposes a new feature based on an aggregation method using the generalized Choquet integral, incorporating a parameter α. To validate its effectiveness, we perform anomaly detection with Random Forest and XGBoost models, assessing its impact on detecting Hulk DoS attacks. Experimental results showed that our proposal significantly improved accuracy. The Random Forest model increased from 93.53% to 97.82% (4.59% improvement), while XGBoost rose from 93.52% to 97.69% (4.46% increase). More importantly, recall for the minority class (attacks) improved substantially, from 0.68 to 0.93 (37% increase) with Random Forest and from 0.68 to 0.92 (35% increase) for XGBoost, reducing false negatives and enhancing intrusion detection. These findings highlight the potential of Choquet-based feature engineering in improving anomaly detection.

Keywords: Anomaly detection · Choquet Integral · Fuzzy · Random Forest · XGBoost

1 Introduction

Computer networks are an integral part of daily life worldwide, enabling global communication and supporting essential services such as banking, healthcare, education, and business operations. In today's interconnected society, the increasing reliance on network infrastructure highlights not only its importance but also its vulnerability [8].

Data network traffic has played numerous roles in the field of security; however, issues such as data leakage, user information exposure, and authentication

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
O. Gervasi et al. (Eds.): ICCSA 2025, LNCS 15650, pp. 3–16, 2025.
https://doi.org/10.1007/978-3-031-96962-1_1

remain significant challenges in the current network environment. Furthermore, the exponential growth of data volume intensifies the complexity of managing and analyzing network traffic, requiring efficient techniques to process and extract relevant information [5].

Among the most critical threats to network stability are Distributed Denial of Service (DDoS) attacks. Given the growing dependency on uninterrupted data services, malicious users may attempt to harm these services to disrupt access for individuals or organizations. In such attacks, multiple computers are hijacked and used to flood the target with excessive traffic, overwhelming the infrastructure and rendering it inaccessible.

These malicious activities often generate identifiable patterns or irregularities, known as anomalies. Detecting such anomalies is essential for mitigating attacks effectively. However, due to the high dimensionality and complexity of modern network data, it becomes necessary to reduce the dimensionality of the data extracted from the network and focus on the most relevant features. Concerned with current network security challenges, this study aims to detect anomalies [16] by implementing a feature engineering approach. The proposed method leverages techniques such as Feature Selection and the Generalized Choquet Integral to enhance the detection and mitigation of DDoS attacks.

In light of these challenges, the Choquet integral serves as a powerful tool for data aggregation, as the fuzzy measure allows for effectively modeling relationships within the given data [2,6]. Despite various methodologies in the literature aimed at enhancing existing algorithms such as Random Forest and XGBoost, no prior studies have applied the generalized Choquet integral to generate features for improving predictive performance. Addressing this gap, this paper introduces a novel feature engineering approach for network traffic analysis, leveraging the fuzzy Choquet integral framework. A key aspect of this method used in this work is the parameter α, which significantly impacts both the execution time and the accuracy of anomaly detection. The effectiveness of this approach is validated using real-world data from a reliable source, demonstrating its practical applicability in network management.

The remainder of the paper is organized as follows. Section 2 presents the related work. The proposal is presented in Sect. 3, and details of its evaluation in Sect. 4. Finally, in Sect. 5, some final remarks are made, and directions for future research are indicated.

2 Related Work

This section discusses research efforts related to the scope of this paper. Studies were gathered that made significant contributions in the areas of DDoS anomaly detection, fuzzy models, and network traffic management.

Wani *et al.* [16] present three well-known machine learning algorithms used in the field of network security, Random Forest, Naïve Bayes, and Support Vector Machine, to detect Distributed Denial of Service (DDoS) attacks. The study aims to contribute to the mitigation of DDoS attacks, which are classified as

critical threats due to their potential to compromise network availability. In this study, the authors used a few tools to set up the testing environment and simulate the attacks. They used OwnCloud, an open-source cloud platform, as the target for the attacks, while the DDoS attacks were carried out using the Tor Hammer tool. Everything was run on the Kali Linux 2018.2 system (Kernel 4.15.0, GNOME 3.28.0). The paper also introduced a dataset with 9 features and 4 class labels, used to evaluate their algorithms. To measure performance, they used metrics like precision, recall, specificity, and F-measure. Overall, the SVM algorithm showed the best results across all metrics, followed by Random Forest.

In addition, Jiang et al. [10] proposed a model for detecting network intrusions using Particle Swarm Optimization (PSO) combined with eXtreme Gradient Boosting (XGBoost). Their goal was to improve accuracy, fine-tune parameters, and select the most relevant features to create a more effective anomaly detection model. Compared to other techniques, the PSO-XGBoost model showed a clear improvement in detecting anomalies. Even with the good results, the authors point out that achieving high precision in Network Intrusion Detection Systems (NIDS) remains a significant challenge, especially when dealing with anomaly detection.

Experimental results demonstrate that the PSO-XGBoost model usually outperforms traditional approaches such as Random Forest, Bagging, and AdaBoost, particularly in detecting minority-class attacks like U2R and R2L. However, the literature lacks studies that explore models for anomaly detection while maintaining low complexity, as fuzzy models do.

Aziz *et al.* [3] proposed a hybrid model to identify the core traffic of the network. First, a payload-based approach identifies most of the traffic. Any remaining unidentified traffic is then analyzed using a statistical unsupervised machine learning method, ensuring that no traffic is left unidentified. The main contributions include, capturing local IP traffic from a university network, identifying traffic using a payload-based method, labeling the unidentified traffic, and applying an unsupervised machine learning approach to classify the rest. In the proposed framework, DPI (Deep Packet Inspection) is employed as the first stage to classify application-layer traffic using predefined protocol signatures. Traffic flows that remain unidentified by DPI are then subjected to machine learning supervised or unsupervised methods. The authors developed and evaluated the framework using real traffic data, and the results demonstrate that the hybrid model effectively addresses the limitations of standalone DPI and machine learning methods. Although the study presents a hybrid system offering a solution for real-time traffic categorization, the proposal lacks a specialized solution for detecting anomalies in network traffic.

Following the networking traffic management field, Shetty *et al.* [15] presented a comprehensive study on the application of Artificial Intelligence (AI) and Machine Learning (ML) for intelligent network traffic control in modern communication systems. The primary objective is to address the limitations of traditional traffic management methods such as static routing, traffic shaping,

and load balancing, which struggle to adapt to the dynamic and high-volume traffic of the today's networks. The authors propose the integration of AI/ML techniques to enhance traffic prediction, congestion control, bandwidth optimization, latency reduction, and Quality of Service (QoS) improvements. Supervised, unsupervised, reinforcement, and deep learning models are analyzed for their effectiveness in managing traffic flows. Real-world case studies from companies like Vodafone, Akamai, and Cloudflare demonstrate the practical value of AI/ML in telecom and content delivery networks.

Yao Hu and Bibo Tu [9] investigated an attack detection method based on the Fuzzy C-Means (FCM) clustering algorithm. Their approach focuses on identifying abnormal network behavior by enabling the system to autonomously learn both typical and atypical service patterns while effectively managing complex network interactions. Once malicious traffic is detected through FCM clustering, the system assesses the severity of the intrusion and initiates appropriate countermeasures. To evaluate the performance of the FCM algorithm, the author used two metrics: miss rate and bit error rate. These results were then compared to those from K-Means and DTSOM. The findings showed that FCM performed better, with lower error rates than the other methods.

As mentioned earlier, these popular machine learning algorithms tend to achieve high accuracy, but they often face challenges when adapting to network behavior, which can lead to false positives and limited generalization. Additionally, although hybrid models can improve accuracy in classifying anomalies, their main focus is usually on categorizing traffic rather than directly detecting anomalies. As a result, some vulnerabilities may remain in the network without being properly addressed.

To address these gaps in the literature, this paper proposes an anomaly detection method that leverages the Choquet integral to create a new feature and enhance detection performance. By incorporating fuzzy measures, the approach improves decision-making. Unlike traditional machine learning models [16], which rely heavily on accurate classifications, this method adaptively weighs traffic features, capturing complex dependencies and improving anomaly detection. As noted in the literature, two models commonly used in similar studies are XGBoost and Random Forest. Therefore, both models will be employed in this work to evaluate and compare the effectiveness of the proposed approach. This method offers a more flexible detection mechanism, capable of recognizing legitimate variations in network behavior and using them to improve detection accuracy.

3 Proposed Approach

This work proposes a novel approach to feature engineering, aiming to enhance anomaly detection regarding DDoS attacks. To achieve this, we first employ KBest [13] as a feature selection algorithm to identify the most relevant features for classification using Random Forest and XGBoost. Once the optimal set of features is determined, we further add a new variable derived from the prediction

of network traffic using a Generalized Choquet Integral, with an adaptive α. This transformation allows the selected feature to be reformulated and reintegrated into the final set of chosen features. By incorporating this new refined feature, the study seeks to improve the model's accuracy, reliability, and efficiency in detecting anomalies.

3.1 Conceptual Model

Figure 1 provides the foundation of this work, detailing its stages and illustrating each step of the process along with its main components and interactions.

1. **Data**: A networking DDoS attack dataset is taken. (Step 1 in Fig. 1)
2. **Data preprocessing**: The data is processed to improve efficiency within the algorithms. (Step 2 in Fig. 1)
3. **Application of feature selection algorithm**: As the data have been processed, a feature selection algorithm is used to choose a range of the best features to use in the prediction algorithm. (Step 3 in Fig. 1)
4. **Application of Choquet integral to aggregation of data from one of the features**: Having selected the features, the first one is taken and the algorithm based on the generalizations of the Choquet integral with adaptive α is used, thus creating a new feature through aggregation and prediction of trends. (Step 4 in Fig. 1)
5. **Integrate the new feature with the selected ones**: This new feature is added along with the previously selected features. (Step 5 in Fig. 1)

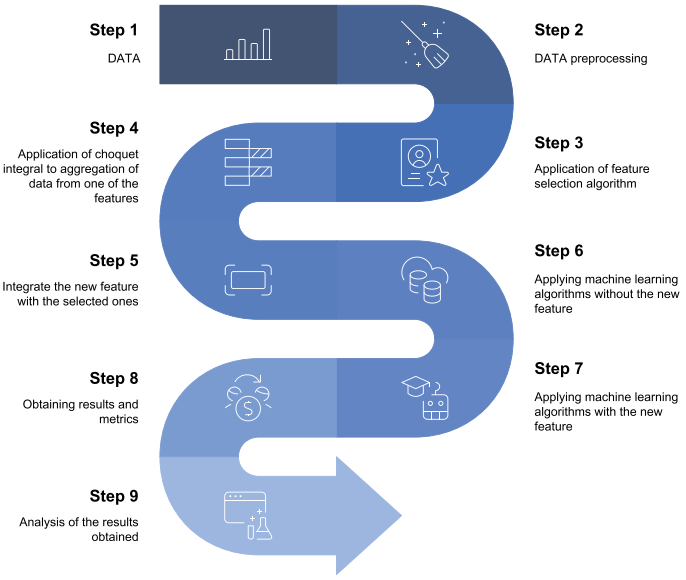


Fig. 1. Conceptual model.

6. **Applying machine learning algorithms without the new feature**: The algorithms used for prediction through machine learning are first applied to the data without the new feature. (Step 6 in Fig. 1)
7. **Applying machine learning algorithms with the new feature**: Then, the same algorithms are applied again, this time with the new feature included. (Step 7 in Fig. 1)
8. **Obtaining results and metrics**: Once the algorithms have been applied, results and performance metrics are obtained. (Step 8 in Fig. 1)
9. **Analysis of the results obtained**: Finally, the results and metrics are analyzed to evaluate the impact of the new feature. (Step 9 in Fig. 1)

3.2 Aggregation and Choquet Integral Functions

The Choquet integral [7] plays a fundamental role in data aggregation [4]. It is defined for a fuzzy measure $m : 2^N \to [0, 1]$, enabling a flexible way to combine inputs by considering the significance of each criterion.

$$\mathcal{C}_m(x) = \sum_{i=1}^{n} \left(x_{\sigma(i)} - x_{\sigma(i-1)}\right) m\left(A_{(i)}\right),$$

where $(x_{\sigma(1)}, \ldots, x_{\sigma(n)})$ represents a permutation of the input values arranged in ascending order, and $A_{(i)} = \{(i), \ldots, (n)\}$ is the subset of indices that correspond to the $n - i + 1$ highest elements of the input x.

Copulas [1] are mathematical functions, $C : [0, 1]^2 \to [0, 1]$, that establish relationships between two-dimensional probability distributions and their respective one-dimensional marginal distributions. They play a key role in the study of probabilistic metric spaces and have extensive applications in statistical analysis.

By leveraging the distributive property of the product, the Choquet integral can be expressed in an expanded form, which is particularly useful when dealing with α-parameterized copulas:

$$\mathcal{C}_m(x) = \sum_{i=1}^{n} \left(x_{(i)} \cdot m\left(A_{(i)}\right) - x_{(i-1)} \cdot m\left(A_{(i)}\right)\right),$$

In [11], this extended representation of the Choquet integral has been further generalized through copula functions, where both product operators are substituted by alternative copula-based methods.

In this work, the copula functions under consideration[1] are summarized in Table 1.

[1] Additional definitions, as well as their properties and mathematical formulations, can be found more deeply in [12].

Table 1. Table of Generalizations of the Choquet integral used to create a new feature.

Copula ID	Functions	α Properties
(A)	$C_\alpha(x,y) = xy[1+\alpha(1-x)(1-y)]$	$-1 \leq \alpha \leq 1(\alpha \neq 0)$
(B)	$C_\alpha(x,y) = \frac{1}{1+\alpha} Max[x+y-1+\alpha-\alpha\|x-y\|,0]$	$0 < \alpha < 1$
(C)	$C_\alpha = (1-\alpha)W + \alpha Min$	$0 < \alpha < 1$
(D)	$C_\alpha = \frac{\alpha^2(1-\alpha)}{2} W + (1-\alpha^2)P + \frac{\alpha^2(1+\alpha)}{2} Min$	$-1 < \alpha < 1(\alpha \neq 0)$

3.3 Dataset

The CIC-DDoS2019 [14] dataset is commonly used in machine learning research focused on network security, particularly for distinguishing between Distributed Denial of Service (DDoS) attacks and normal traffic. It was developed by the Canadian Institute for Cybersecurity (CIC) and the data was collected using CICFlowMeter-V3 that includes traffic from multiple devices, providing a scenario closer to a real network environment.

The dataset was compiled by CIC and is part of their CIC-DDoS2019 collection. It contains a mix of normal traffic and DDoS-related packet data. With over 200,000 rows, it provides data to train and evaluate machine learning models. The dataset includes a variety of traffic features, making it suitable for feature engineering and model optimization. The HULK attack, present in the CIC-DDoS2019 dataset, is a type of DoS attack that floods servers with a high volume of HTTP requests to make them unavailable. This attack was used to evaluate our approach in this work.

3.4 Implementation

For the implementation of this work, we used four algorithms: Random Forest, SelectKBest, XGBoost, and the Choquet Integral - the first three being well-known and widely used in the academic community.

It is worth noticing that SelectKBest is a machine learning technique used for feature selection, and it was applied in this work to identify and choose the most relevant variables from the dataset. This approach enhances model performance while reducing the risk of overfitting. SelectKBest is one of the most widely used methods within this technique and falls under the category of filter-based feature selection. It applies statistical tests such as Chi-Squared, ANOVA F-test, and Mutual Information Score to rank features based on their relationship with the target variable, selecting the top K features with the highest scores.

To predict DoS attacks, we employed two machine learning algorithms: Random Forest and XGBoost. Although both are based on decision trees, they utilize distinct techniques to enhance accuracy and robustness. Random Forest is an ensemble learning method that constructs multiple decision trees from different subsets of the training data. By selecting random features at each split, it reduces overfitting and improves generalization. The final prediction is determined by aggregating the outputs of all trees—typically through majority voting

in classification tasks. This approach not only enhances robustness to noise but also provides valuable internal estimates for error monitoring. On the other hand, XGBoost (eXtreme Gradient Boosting) is a highly efficient and scalable gradient-boosting algorithm. Unlike Random Forest, it builds decision trees sequentially, with each new tree correcting the errors of the previous ones using gradient descent optimization. XGBoost is well known for its speed and ability to handle large datasets. We can have a look at how it behaves in the Algorithm 1.

Algorithm 1. Feature Selection and Model Evaluation

Input: Dataset D
Output: Model performance metrics

$D_{clean} \leftarrow$ Preprocess(D) ▷ Data cleaning and preprocessing
$selectedFeatures \leftarrow$ KBest(D_{clean}, $k = 4$) ▷ Select top 4 features using KBest
$choquetFeature \leftarrow$ ChoquetIntegral($selectedFeatures[0]$) ▷ Generate additional feature
$finalFeatures \leftarrow selectedFeatures \cup \{choquetFeature\}$ ▷ Combine features
$D_{final} \leftarrow$ PrepareData(D_{clean}, $finalFeatures$) ▷ Prepare dataset with selected features
$model1 \leftarrow$ Train(RandomForest, D_{final}) ▷ Train Random Forest
$model2 \leftarrow$ Train(XGBoost, D_{final}) ▷ Train XGBoost
$results1 \leftarrow$ Evaluate($model1$, D_{final}) ▷ Evaluate Random Forest
$results2 \leftarrow$ Evaluate($model2$, D_{final}) ▷ Evaluate XGBoost
$model1_{noFeature} \leftarrow$ Train(RandomForest, $D_{noFeature}$) ▷ Train Random Forest without new feature
$model2_{noFeature} \leftarrow$ Train(XGBoost, $D_{noFeature}$) ▷ Train XGBoost without new feature
$results1_{noFeature} \leftarrow$ Evaluate($model1_{noFeature}$, $D_{noFeature}$) ▷ Evaluate Random Forest without new feature
$results2_{noFeature} \leftarrow$ Evaluate($model2_{noFeature}$, $D_{noFeature}$) ▷ Evaluate XGBoost without new feature
return $results1$, $results2$, $results1_{noFeature}$, $results2_{noFeature}$ ▷ Return model performance metrics

The Choquet-based algorithm employed in this work was developed using copulas, as presented in Table 1, it is used for data aggregation and trend prediction. The algorithm takes a selected feature as input and processes it using specific techniques to compute the mean absolute error (MAE) associated with each of the four copulas and its α related to them. The copula and the α with the lowest MAE are then selected to produce the final output which will be used in the classification algorithms.

4 Evaluation

To prove the technical feasibility of the proposal, this section evaluates the influence of using the Generalization of the Choquet integral as a feature creator, and its results when used in algorithms to detect anomalies.

4.1 Generalization of the Choquet Integral

Was address the implementation of a prediction model using the Choquet Integral applied to a feature extracted from the dataset, where the mean absolute error (MAE) analysis is performed with different copulas (Table 1) and α values associated. Its possible to observe an algorithm example in (Algorithm 2) of how the best α of each copula was identified, the parameter such as upper and lower bounds to get the most useful α was selected after an analysis of the metrics from the MAE combined with number of interactions, was concluded that after one hundred interactions the continuation of the search algorithm becomes irrelevant, returning a minimally acceptable improvement values.

Algorithm 2. Binary Search Algorithm to Optimize α for Choquet Integral

Input: $\alpha_{\min}, \alpha_{\max}$ (initial bounds of the α interval), $f(\alpha)$ (error function), max_iterations (maximum iterations), tolerance (convergence tolerance)
Output: α_{opt} (optimized value of α)

$iteration \leftarrow 0$ ▷ Initialize iteration counter
$best_error \leftarrow \infty$ ▷ Initialize best error as infinity
while $iteration <$ max_iterations **do**
  $\alpha \leftarrow \frac{\alpha_{\min}+\alpha_{\max}}{2}$ ▷ Compute midpoint of interval
  $error \leftarrow f(\alpha)$ ▷ Evaluate error function at α
  if $error < best_error$ **then**
    $best_error \leftarrow error$ ▷ Update best error found
    $\alpha_{\text{opt}} \leftarrow \alpha$ ▷ Store best α value
  end if
  if $error >$ tolerance **then**
    $\alpha_{\min} \leftarrow \alpha$ ▷ Adjust lower bound of search interval
  else
    $\alpha_{\max} \leftarrow \alpha$ ▷ Adjust upper bound of search interval
  end if
  $iteration \leftarrow iteration + 1$ ▷ Increment iteration counter
end while
return α_{opt} ▷ Return optimized α

First, when running the Choquet algorithm, we computed the tendencies of each previously imputed feature value for each copula. Next, we applied a binary search to identify the optimal α value for each copula. The lowest error was found after 8 iterations of the binary search algorithm. The best result for our dataset was achieved with $\alpha = 0.5342$, using the first copula (Table 1) in the generalized Choquet integral.

4.2 Data Preprocessing and Feature Selection Algorithm

Before selecting and generating new features, several preprocessing steps were performed as part of this study to ensure data quality and improve model performance.

Since we started with the raw CIC-DDoS 2019 dataset, data preprocessing was necessary. This process was carried out using the Pandas and NumPy libraries for data manipulation and cleaning. The preprocessing steps included converting categorical columns into numerical values using techniques such as one-hot encoding and label encoding, handling missing data through mean/mode imputation, and balancing the dataset with undersampling and SMOTE (Synthetic Minority Over-sampling Technique) to address class imbalances.

Once the data was preprocessed, the SelectKBest algorithm from the Scikit-learn library was applied with a parameter value of 4, using the ANOVA F test, a scoring function to rank and select the four most relevant features for prediction. This step helped reduce dimensionality while retaining the most informative attributes, as there were 68 features previously and then 4 features were selected with the highest scores, which we can see listed below along with their respective scores:

- Bwd Packet Length Max (112023.056)
- Bwd Packet Length Mean (112023.056)
- Bwd Packet Length Std (108310.460)
- Avg Bwd Segment Size (106125.975)

After selecting these features, a copy of the first one was extracted and used as input for the generalized Choquet integral function, leveraging the fuzzy measure-based aggregation technique to compute an additional feature. This new feature captures non-linear interactions between the selected attributes, potentially improving predictive performance.

Finally, a new dataset was constructed, containing only the four selected features, the newly generated Choquet feature, and the label column. This refined dataset was then used to train and evaluate machine learning models, ensuring a more efficient and robust classification process.

4.3 Application of Random Forest and XGBoost Algorithms on Data

After preprocessing the data, selecting new features, and generating the additional feature, the dataset was ready for the prediction phase. To carry out the predictions, two algorithms explained in Subsect. 3.4 were used: Random Forest and XGBoost. For each algorithm, two tests were conducted: the first without the newly generated feature and the second with its inclusion. This approach provided insights into the model's improvement after incorporating the new feature.

It is important to highlight the parameters used for both algorithms, Random Forest and XGBoost, you can see it in Table 2 and Table 3 respectively:

Table 2. Hyperparameters used for Random Forest

Parameter	Value	Description
`n_estimators`	100	Number of trees in the forest
`max_depth`	10	Maximum depth of each tree
`random_state`	42	Ensures reproducibility
`stratify`	y	Maintains class distribution in train/test split

Table 3. Hyperparameters used for XGBoost

Parameter	Value	Description
`n_estimators`	100	Number of boosting rounds
`max_depth`	6	Maximum depth of each tree
`learning_rate`	0.1	Step size for each boosting iteration
`subsample`	0.8	Percentage of data used per boosting round
`colsample_bytree`	0.8	Percentage of features used per tree
`random_state`	42	Ensures reproducibility
`use_label_encoder`	False	Avoids deprecated label encoding warnings
`eval_metric`	logloss	Loss function for evaluation

4.4 Discussion of Results

Through the data and tables presented below, we will discuss the results obtained in this work.

Feature Selector. Using the SelectKBest algorithm, only four features were selected from the numerous ones that composed the initial dataset: Bwd Packet Length Max, Bwd Packet Length Mean, Bwd Packet Length Std, and Avg Bwd Segment Size. Among these, Bwd Packet Length Max was chosen as the parameter for creating a new feature using the Choquet Integral, this selection was based on the SelectKBest ranking recommendation, taking the first ranked of these four features to pass throw the Choquet algorithm.

Random Forest. Without the new feature, the model achieved an accuracy of 93.53%, with a recall of 0.68 for the minority class (1), indicating difficulties in detecting attacks. After incorporating the new feature, accuracy increased to 97.82%, with recall for class 1 rising to 0.93, reflecting a significant improvement in attack detection. Additionally, the F1-score for the minority class improved from 0.81 to 0.94, demonstrating that the model became more effective at identifying attacks without compromising precision in classifying the majority class. For a more detailed analysis, one can refer to the Table 4.

Table 4. Performance Metrics for Random Forest

Metric	Without New Feature	With New Feature	Improvement %
Accuracy	0.9353	0.9782	4.59%
Recall (Class 1)	0.68	0.93	37%
F1-score (Class 1)	0.81	0.94	16%

XGBoost. Similar to Random Forest, XGBoost initially achieved an accuracy of 93.52%, with a recall of 0.68 for the minority class. After adding the new feature, accuracy increased to 97.69%, and recall for class 1 improved to 0.92. The F1-score also saw a significant boost, rising from 0.81 to 0.94 (Table 5).

Table 5. Performance Metrics for XGBoost

Metric	Without New Feature	With New Feature	Improvement %
Accuracy	0.9352	0.9769	4.46%
Recall (Class 1)	0.68	0.92	35%
F1-score (Class 1)	0.81	0.94	16%

5 Conclusion

This work introduces a new feature based on the Generalization of the Choquet Integral to enhance anomaly detection using Random Forest and XGBoost models. The evaluation presented in this work showed a significant improvement in detecting the HULK attack after adding the new feature. A general improvement was also observed, but the most notable gains were in recall and F1 score for the minority class (1). This indicates that the new feature reduced false negatives, making the model more effective at detecting anomalies.

The analysis of the results shows that introducing this new feature helps machine learning algorithms better identify complex patterns in the data. This finding highlights the promise of the proposed approach to improve the detection of intrusions and anomalies in network environments. Looking ahead, the team plans to apply the Choquet Integral to different datasets and assess how well it performs in detecting various types of DDoS attacks.

Acknowledgment. The authors would like to thank Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS) (24/2551-0001396-2) and FAPERGS/CNPq (23/2551-0000126-8; 23/2551-0000773-8).

References

1. Alsina, C., Schweizer, B., Frank, M.J.: Associative functions: triangular norms and copulas. World Scientific (2006)
2. Amorim, M., et al.: Systematic review of aggregation functions applied to image edge detection. Axioms **12**(4) (2023). https://doi.org/10.3390/axioms12040330, https://www.mdpi.com/2075-1680/12/4/330
3. Aziz, W.A., Qureshi, H.K., Iqbal, A., Al-Dulaimi, A., Al–Rubaye, S.: Towards accurate categorization of network ip traffic using deep packet inspection and machine learning. In: GLOBECOM 2023 - 2023 IEEE Global Communications Conference, pp. 01–06 (2023). https://doi.org/10.1109/GLOBECOM54140.2023.10437078
4. Beliakov, G., Pradera, A., Calvo, T.: Aggregation functions: A guide for practitioners, vol. 221. Springer (2007)
5. Cardoso, F.C., Berri, R.A., Borges, E.N., Dalmazo, B.L., Lucca, G., de Mattos, V.: Echo state network and classical statistical techniques for time series forecasting: a review. Knowl.-Based Syst. **293**, 111639 (2024). https://doi.org/10.1016/j.knosys.2024.111639, https://www.sciencedirect.com/science/article/pii/S0950705124002740
6. Carpenter, J., Layne, J., Serra, E., Cuzzocrea, A., Gallo, C.: Structural node representation learning for detecting botnet nodes. In: Gervasi, O., et al. ,(eds.) Computational Science and Its Applications – ICCSA 2023, pp. 731–743. Springer Nature Switzerland (2023). https://doi.org/10.1007/978-3-031-36805-9_47
7. Choquet, G.: Theory of capacities. Annales de l'Institut Fourier **5**, 131–295 (1953)
8. Dalmazo, B.L., Vilela, J.P., Curado, M.: Triple-similarity mechanism for alarm management in the cloud. Comput. Secur. - Elsevier **78**, 33–42 (2018). https://doi.org/10.1016/j.cose.2018.05.016, http://www.sciencedirect.com/science/article/pii/S0167404818306515
9. Hu, Y., Tu, B.: Security situation assessment model of ddos attack based on progressive fuzzy c clustering algorithm. In: 2024 International Conference on Data Science and Network Security (ICDSNS), pp. 1–4 (2024). https://doi.org/10.1109/ICDSNS62112.2024.10691183
10. Jiang, H., He, Z., Ye, G., Zhang, H.: Network intrusion detection based on PSO-Xgboost model. IEEE Access **8**, 58392–58401 (2020). https://doi.org/10.1109/ACCESS.2020.2982418
11. Lucca, G., Dimuro, G.P., Bedregal, B., Sanz, J.A., Bustince, H.: A proposal for tuning the α parameter in $c_\alpha c$-integrals for application in fuzzy rule-based classification systems. Nat. Comput. **19**(3), 533–546 (2020)
12. Lucca, G., Dimuro, G.P., Bedregal, B., Sanz, J.A., Bustince, H.: A proposal for tuning the alpha parameter in a copula function applied in fuzzy rule-based classification systems. In: 2016 5th Brazilian Conference on Intelligent Systems (BRACIS), pp. 367–372 (2016). https://doi.org/10.1109/BRACIS.2016.073
13. Otchere, D.A., Ganat, T.O.A., Ojero, J.O., Tackie-Otoo, B.N., Taki, M.Y.: Application of gradient boosting regression model for the evaluation of feature selection techniques in improving reservoir characterisation predictions. J. Petroleum Sci. Eng. **208**, 109244 (2022). https://doi.org/10.1016/j.petrol.2021.109244, https://www.sciencedirect.com/science/article/pii/S0920410521008998
14. Sharafaldin, I., Lashkari, A.H., Hakak, S., Ghorbani, A.A.: Developing realistic distributed denial of service (DDoS) attack dataset and taxonomy. In: 2019 International Carnahan Conference on Security Technology (ICCST), pp. 1–8 (2019). https://doi.org/10.1109/CCST.2019.8888419

15. Shetty, S., S, T.M., M, V.H., Shaikh, R.N.: Intelligent network traffic control with AI and machine learning. In: 2024 IEEE 16th International Conference on Computational Intelligence and Communication Networks (CICN), pp. 353–357 (2024). https://doi.org/10.1109/CICN63059.2024.10847397
16. Wani, A.R., Rana, Q.P., Saxena, U., Pandey, N.: Analysis and detection of DDoS attacks on cloud computing environment using machine learning techniques. In: Proceedings of the 2019 Amity International Conference on Artificial Intelligence (AICAI), pp. 870–875 (2019). https://doi.org/10.1109/AICAI.2019.8701238