ERPCA-GFCM: An Enhanced DDoS Detection Model on Internet of Things

Grace Anne S. Cahulogan [1], Ederlyne Ann V. Gordula [2], Vivien A. Agustin [3], Herminiño C. Lagunzad [4]

[1] BS Computer Science Student, Computer Science Department, Pamantasan ng Lungsod ng Maynila, Philippines, gascahulogan2019@plm.edu.ph

[2] BS Computer Science Student, Computer Science Student, Pamantasan ng Lungsod ng Maynila, Philippines, eavgordula2019@plm.edu.ph

[4] Thesis Adviser, Computer Science Department, Pamantasan ng Lungsod ng Maynila, Philippines, vaagustin@plm.edu.ph

[5] Thesis Adviser, Computer Science Department, Pamantasan ng Lungsod ng Maynila, Philippines, hclagunzad@plm.edu.ph

Abstract: This research paper introduced an Enhanced Distributed Denial-of-Service (DDoS) Detection model specifically designed for IoT devices. Given the prevalence of DDoS attacks targeting IoT devices, which involve overwhelming a system with malicious traffic to disrupt its normal functioning, the proposed model aimed to enhance the security and resilience of IoT networks. To address this, the proposed model integrated multiple techniques to improve detection and classification accuracy. The first technique, ER-Relief algorithm, is a feature selection method made to address the presence of noise and outliers in the dataset by minimizing a loss function based on the empirical sum of margins. To further enhance the model's performance, Principal Component Analysis (PCA) was utilized for dimensionality reduction. PCA transforms the original high-dimensional feature space into a lower-dimensional space while preserving the most critical information. To achieve better clustering results, the model incorporates the Global Fuzzy C-means algorithm. This algorithm addressed the issue of sensitivity to initial conditions, which lead to suboptimal clustering results. By incorporating fuzzy logic principles, Global Fuzzy C-means assigning data points to multiple clusters with varying degrees of membership, providing a more nuanced representation of the underlying data structure. Lastly, the Random Forest algorithm was employed for training and testing the model. The model was then tested on the CICDDoS2019 dataset, which contains three (3) types of DDoS attacks, namely DNS, UDP, and MSSQL attacks. Based on the evaluation results, the proposed model achieved an impressive accuracy of 97.92%, recall of 97.92%, F1-score of 97.90%, and precision of 97.93%. These metrics highlighted the model's effectiveness, showcasing its ability to accurately detect and classify various types of DDoS attacks with high precision and recall. This research contributes to the advancement of network security by providing a robust and reliable solution for combating DDoS attacks.

Keywords: fuzzy c-means, er-relief, principal component analysis, random forest, internet of things, ddos detection.

1. INTRODUCTION

A. Background of the Study

The continuous rise in popularity and growth of the Internet of Things (IoT) has unquestionably made IoT devices a prime target for cyber security attacks. Reference [9] shows that it is evident that these devices are highly susceptible to security incidents, making them the weakest link in modern computer networks. However, ensuring the stability and dependability of IoT devices is a complex task due to their highly distributed and interconnected network structure [15].

One common disruption to a network's accessibility is Distributed Denial-of-Service (DDoS) attacks, which are known for their high intensity and low cost to initiate. These attacks can cause immediate and severe harm to the targeted

victims. DDoS attacks are typically carried out using a network of compromised devices called a botnet, which floods the targeted system with requests, potentially overwhelming it and rendering it unable to respond to normal traffic (denial-of-service).

Implementing FCM comes with drawbacks. FCM can be sensitive to initial conditions, resulting in different clustering outcomes with varying initializations. It is also sensitive to noise and outliers, affecting the accuracy of clustering. Moreover, FCM may face challenges in handling high-dimensional data due to the curse of dimensionality, making clustering complex.

The researchers proposed modifications to enhance the FCM's DDoS detection capability. They used the CICDDoS2019 dataset, which contains real-world packet capture data of benign traffic and various DDoS attacks like DNS, UDP, and MSSQL attacks. To handle noise and outliers, they employed ER-Relief, a combination of Exact-Relief and Ramp-Relief algorithms, for feature selection. Principal component analysis (PCA) was used to reduce data dimensionality and improve model performance. They implemented Global Fuzzy C-Means (GFCM) to minimize sensitivity to initial conditions by running multiple iterations with different initializations and combining the results. The proposed model was evaluated using metrics such as confusion matrix, accuracy, precision, F1-Score, and recall, demonstrating its superior accuracy and consistency compared to traditional FCM.

B. Statement of the Problem

The researchers aim to propose modifications to enhance its effectiveness in detecting DDoS attacks on IoT devices.

1. *Sensitivity to Initial Conditions.* The FCM is sensitive to initial conditions and can produce local minimum outcomes. It heavily relies on the manual initialization of the membership matrix and the number of clusters [22]. Additionally, due to the non-convex nature of the Fuzzy clustering targets, starting with an incorrect initial condition may lead the algorithm to converge on a local extremum instead of the optimal solution [12].
2. *Sensitivity to noise and outliers*. According to Reference [3], the FCM algorithm, which is a widely used fuzzy clustering algorithm, is very sensitive to noise, outliers, and the size of clusters.
3. *Problem in dealing with high dimensional data.* References [33] and [33] specified that the fuzzy c-means clustering algorithm encounters difficulties when working. with high-dimensional data. This observation was later corroborated by [7], who also noted the challenges of FCM when dealing with high-dimensional data.

C. Objective of the Study

The study specifically focuses on addressing the following issues:

1. To solve Fuzzy C-means' sensitivity to initial conditions by using an enhanced version of the algorithm called Global Fuzzy C-means clustering to help detect DDoS attacks in IoT devices.
2. To achieve a solution to Fuzzy C-means' sensitivity to noise and outliers by using ER-Relief as a feature selection method.
3. To produce low-dimensional data by using Principal Component Analysis (PCA) to reduce dimensionality.

D. Significance of the Study

By showcasing the effectiveness of the modified algorithm, the study could contribute to the following:
Users. including employees using IoT devices at work or individuals with IoT devices at home. It provides knowledge about the severity of DDoS attacks and their impact on IoT devices.
Security Companies. This study assists companies and industries in understanding how to combat DDoS attacks, protect assets, and prevent economic consequences. It also aids in developing stronger methods for safeguarding IoT devices, improving overall security and stability.
Future Researchers. This study can also provide Future Researchers recommendations and other factors that they can consider in doing future enhancements.

E. Scope and Limitations

This study focused on addressing issues of the FCM algorithm, including sensitivity to initial conditions, noise and outliers, and the high dimensionality of data. The research specifically targeted IoT devices and narrowed down the dataset to 3202 network flow data involving three types of DDoS attacks: DNS and SQL Injection (reflection-based) and UDP flood (exploitation-based). The dataset also included normal or benign traffic for model evaluation. No other dataset was used in this study.

F. Definition of terms

Botnets - a network of interconnected devices infected with malware and are controlled without permission.
Clustering - is a machine learning method to group data points. In this study, it is used to detect DDoS attacks among normal flows.
CICDDoS2019 - the DDoS dataset, consists of benign and latest variants of DDoS attacks, used by the researchers in this study which resembles real-world packet capture (PCAPs).
Dimensionality Reduction - the process of reducing features, as well as dimensions in the data, particularly for this study, the CICIDDoS2019 dataset. This method helps in simplifying the representation of data while keeping significant information.
Distributed Denial of Service (DDoS) - the main application in this study and is detected using Fuzzy C-means algorithm.
ER-Relief - A relief-based algorithm that is a combination of Exact Relief and Ramp-Relief algorithms, proposed by
Feature Selection - the process of choosing the most important features from a dataset to improve model performance.
Global Fuzzy C-means Algorithm - the clustering algorithm used in this study to detect DDoS attack in IoT devices represented by CIC-DDoS2019 dataset.
Internet of Things (IoT) - set of interrelated devices where CIC-DDoS2019 that is used in this study originated.
Machine Learning - the practice of teaching computers to learn and make predictions based on data.
Noise - refers to data errors, random variations, and inconsistencies present and identified in the CIC-DDoS2019 dataset. Due to its ability to affect the reliability and accuracy of the model, data noises will be solved in this study through feature selection using the ER-Relief algorithm.
Outliers - data points that stand out from the rest of the dataset due to their extreme values.
Principal Component Analysis PCA) - A method for dimensionality reduction used in this study.

2. REVIEW OF RELATED LITERATURE

A. Foreign Literature

The number of worldwide DDoS attacks has significantly increased, with a 90% growth in Q3 of 2022 [9] compared to the previous year. These attacks have become more powerful due to the development of new, high-powered DDoS tools by hackers. APAC, particularly Taiwan and the Philippines, experienced the highest percentage of DDoS attacks in 2021, making it a highly targeted region. [23]

Previously, tools like Intrusion Detection Systems (IDS) and firewalls were sufficient for detecting and mitigating DDoS attacks by tracing and blocking the source [17]. However, it is now evident that developing new methods and enhancing existing ones is crucial to cope with the growing scale of DDoS attacks.

There are already three methods in detecting DDoS attacks [13]. The first one is based on protocol analysis, the second one is based on cluster, and the third method is based on network traffic statistics. Protocol analysis-based DDoS attack detection is only effective for attacks with obvious abnormal protocol characteristics and not for those without such characteristics. Statistical methods for DDoS attack detection have limitations, including their reliance on prior knowledge of network flow patterns. However, the challenge arises from the dynamic nature of network flow, making it difficult to accurately characterize. [19].

Clustering methods do not rely on prior knowledge of data distribution [14]. Reference [28] conducted a study where they ranked features from internet traffic datasets and evaluated various machine learning algorithms for DDoS attack detection. FCM stood out as it demonstrated high accuracy and efficiency.

B. Foreign Studies

Reference [16] proposed a new machine learning technique for detecting DDoS attacks in industrial IoT devices. They used graph theory to extract features from traffic data and applied principal component analysis (PCA) to extract additional features. The fuzzy C-means algorithm (FCM) was then used to detect DDoS attacks. The model demonstrated high reliability with a recall of 100.00%, false

positive rate of 1.05%, true positive rate of 68.95%, true negative rate of 0.00%, and false negative rate of 30%.

A method combining supervised and unsupervised algorithms was developed to detect unknown malicious attacks. By using the DBSCAN clustering algorithm and statistical measures, they achieved a significantly higher Positive Likelihood Ratio (LR+) of approximately 198% compared to other machine learning algorithms. The method was evaluated using the CIC-DDoS2019 dataset [21].

Reference [28] developed a model using machine learning and data mining to detect DDoS attacks. They applied different algorithms to the CIC-DDoS2019 dataset and found that AdaBoost and XGBoost were the most accurate, achieving 100% accuracy and an F1-Score of 1. XGBoost also had faster training and detection times compared to AdaBoost. Reference [6] developed a method using semi-supervised fuzzy c-means to detect DDoS botnets. The technique involves learning features indicative of botnet attacks and analyzing network traffic to classify and locate infected hosts.

In a study by [24], a fog-based detection framework was introduced that utilizes fog computing and the ESFCM method. This framework enables distributed detection at the network edge and improves detection for IoT devices. Evaluation using the NSL-KDD dataset demonstrated superior performance with an 11ms detection time and 86.53% accuracy rate compared to centralized frameworks.

C. Synthesis

DDoS attacks pose a significant threat to industries and organizations that have public networks, as they can disrupt web services and result in economic consequences. The Internet of Things (IoT) devices are also vulnerable to such attacks [18]. Therefore, the detection of DDoS attacks is an important area of research.

This study aims to address the issue of the fuzzy c-means algorithm's sensitivity to initial conditions, such as cluster center values and membership values, which can impact the clustering results [27]. The proposed solution involves the implementation of global fuzzy c-means in conjunction with a self-organizing map.

To handle the sensitivity of fuzzy c-means algorithms to noise and outliers in complex and high-dimensional data, this study proposes a modification using the ER-Relief algorithm [32], specifically designed to address this challenge.

This study aims to overcome the challenges of Fuzzy c-means with high-dimensional data by employing principal component analysis (PCA) for feature selection and dimensionality reduction. The goal is to develop a detection model that effectively identifies DDoS attacks using the CIC-DDoS2019 dataset, leveraging enhanced algorithms.

D. Comparative Analysis

Table 1: Comparison of Fuzzy C-means with other Clustering Algorithms

DDoS Detection Systems (Devi et al, 2016)	**Limitations**
DDoS attacks detection based on Protocol Analysis	Effective against attacks with clear abnormal protocol features but less effective against attacks lacking distinct abnormal characteristics.
DDoS attacks detection based on cluster	Requires extensive data and has a high error rate. Additionally, it faces challenges in distinguishing between user visits and actual DDoS attacks in network traffic.

DDoS attacks detection based on network traffic statistics	Creates a model to detect abnormal traffic statistics but struggles to handle such traffic and differentiate between legitimate high traffic and actual DDoS attacks.

In the comparison table, Fuzzy C-means and K-means clustering algorithms are compared. Fuzzy C-means has shown better accuracy in DDoS detection models compared to other algorithms, as observed in the study by Anitha and Suresh (2011). The main difference between the two algorithms is their approach to data. While K-means assigns each data point to a single cluster, Fuzzy C-means allows a data point to belong to multiple clusters based on its degree of membership. Fuzzy C-means, being a well-known algorithm, is chosen for this model.

To improve the Fuzzy C-means (FCM) algorithm and overcome its vulnerabilities, this study implements Global Fuzzy C-means (GFCM). GFCM is a gradual approach that eliminates the sensitivity to initial conditions and the tendency to fall into local minima. By conducting a sequence of local searches, GFCM reduces clustering errors and achieves an optimal solution, addressing a key problem of FCM. The study compares and demonstrates the effectiveness of both FCM and GFCM algorithms.

3. THEORETICAL FRAMEWORK

This chapter introduces key theories, concepts, and algorithms that are essential to the study. It outlines their relevance to the study's objectives and explains how they will be applied in the DDoS detection model.

A. Fuzzy C-Means Algorithm

Zadeh's work on fuzzy sets (1965) aimed to enhance clustering by considering the probability of each parameter. Building upon this concept, Reference [5] developed the fuzzy c-means algorithm, which is widely used in this field.

The concept of a fuzzy c-partition allows for flexible membership of data in multiple classes, unlike a crisp partition where data belongs to only one class. The Fuzzy C-means (FCM) algorithm implements fuzzy partitioning, assigning data membership grades ranging from 0 to 1 in multiple groups. This is beneficial when crisp partitions are difficult to determine (Yuan et. Al, 1995).

This approach assigns membership to data points based on their distance to cluster centers. Closer points have higher membership. The sum of memberships for each point is one. Memberships and cluster centers are updated using a formula after each iteration.

B. ER-Relief Algorithm

The Relief feature selection algorithm is widely recognized [32]. However, it has limitations in handling outliers and noisy features, which are important factors in feature selection.

To overcome these limitations, the researchers introduced the ER-Relief algorithm, which combines both R-Relief and E-Relief. R-Relief focuses on identifying relevant features by comparing instances with similar and dissimilar class labels, while E-Relief addresses noisy features by comparing attribute values within instances.

The objective of the ER-Relief algorithm is to find the best feature weights by maximizing the difference between relief scores obtained from R-Relief and E-Relief. The algorithm ensures that the weights are non-negative and have a unit norm, promoting positivity and continuity during the feature selection process.

The ER-Relief algorithm effectively determines the weights of features by comparing their values and the distances between instances. It assesses the discriminative power of each feature by examining instances with different labels. The algorithm selects the most relevant features by updating and normalizing the weights based on the number of hits and misses. In conclusion, the ER-Relief algorithm is a valuable tool for feature selection in machine learning applications.

C. Principal Component Analysis (PCA)

Principal Component Analysis (PCA), also known as the discrete Karhunen-Loève transform (KLT), the Hotelling transform singular value decomposition (SVD), and empirical orthogonal function (EOF), is an unsupervised dimensionality reduction technique [4]. Without significant loss of essential information from the original data, PCA utilizes the relationships among input features to transform high-dimensional data into a lower-dimensional representation. This enables a more manageable and concise portrayal of the data.

The PCA algorithm seeks to minimize data dimensionality by transforming it into a new set of variables known as principal components. These components are linear combinations of the original features that capture the greatest amount of turbulence in the data. The technique is used in a variety of fields, including image compression, face recognition, pattern recognition, eigenfaces, text categorization, and computer vision [4].

D. Global Fuzzy C-means

The FCM algorithm is a clustering method that relies on local search to find the best clusters based on a given criterion. However, it is sensitive to initial conditions and can converge to a local minimum. To address this, reference [29] introduced a method that involves using the FCM algorithm as a local search approach with multiple initial positions for the cluster centers.

Unlike most global clustering algorithms that arbitrarily choose initial values for all cluster centers, the GFCM technique adds one new cluster center at each stage in an optimal manner. The algorithm's main advantage is its independence from initial conditions, which leads to improved clustering accuracy.

Because the global FCM clustering algorithm does not rely on any initial conditions, it avoids the sensitivity to initial value and enhances clustering accuracy [29].

E. Random Forest Algorithm

The Random Forest Algorithm [7] is a machine learning technique that improves prediction accuracy without significantly increasing computational demands. It constructs a forest of decision trees using a randomized approach. Each decision tree in the forest operates independently, making its own decisions when presented with a new input sample. The collective decisions of the decision trees are used to evaluate and classify the sample.

In a random forest, each classification tree is built as a binary tree using recursive splitting. The tree starts with the root node containing all training data and recursively splits it into left and right nodes based on impurity minimization. This process continues until the branch growth stopping rule is satisfied.

4. RESULTS AND DISCUSSION

The clustering and accuracy results of Global Fuzzy C-means and Random Forest applied on the CIC-DDoS2019 dataset, were assessed using different performance metrics and were also compared to the baseline algorithm, along with variations of the model.

A. Clustering Performance Evaluation Metrics

To know the effectiveness of the clustering of the improved model compared to the baseline algorithm, the two models were also compared to other model variations progressively. Specifically, the models that were compared with each other are Fuzzy C-means (FCM), Fuzzy C-means with ER-Relief (ER-FCM), Fuzzy C-means with ER-Relief and Principal Component Analysis (ER-PCA-FCM), and Global Fuzzy C-means with ER-Relief and Principal Component Analysis (ERPCA-GFCM).

Using the dataset CIC-DDoS2019, the following results were acquired after a series of enhancements. In addition, this section will also feature the clustering performance of each model after ten (10) tries, to test its consistency and accuracy level.

Table 2: Clustering Performance Evaluation of FCM

FCM					
	Silhouette	Adjusted Rand Index	Homogeneity	Completeness	V-measure
1	-0.1575	0.2187	0.2583	0.4852	0.3372
2	0.4176	0.2186	0.2570	0.4902	0.3372
3	-0.1587	0.2183	0.2571	0.4856	0.3369
4	0.4176	0.2186	0.2570	0.4902	0.3372
5	-0.1131	0.2093	0.2621	0.4543	0.3326
6	-0.1630	0.2205	0.2586	0.4883	0.3382
7	0.4180	0.2182	0.2568	0.4898	0.3369
8	0.4180	0.2182	0.2568	0.4898	0.3369
9	0.4176	0.2186	0.2570	0.4902	0.3372
10	-0.1575	0.2187	0.2583	0.4852	0.3372

In Table 2, the results were obviously not consistent as it comprises some reasonable scores, and some that are too low to be considered as well-clustered scores (-0.1575 and -0.1630). On the other hand, the ARI and homogeneity scores showed moderate results in terms of similarity between clusters and degree of membership of each data point per cluster. However, it is not considered as high enough and can be further improved. There is also room for improvement in the completeness and v-measure scores.

Table 3: Clustering Performance Evaluation of ER-FCM

ER-FCM					
	Silhouette	Adjusted Rand Index	Homogeneity	Completeness	V-measure
1	0.7861	0.2544	0.2427	0.6313	0.3506
2	0.7865	0.2541	0.2426	0.6316	0.3506
3	0.7861	0.2544	0.2427	0.6313	0.3506
4	0.7861	0.2544	0.2427	0.6313	0.3506
5	0.7850	0.2553	0.2430	0.6305	0.3508
6	0.7865	0.2541	0.2426	0.6316	0.3506
7	0.7865	0.2541	0.2426	0.6316	0.3506
8	0.7861	0.2544	0.2427	0.6313	0.3506

9	0.7865	0.2541	0.2426	0.6316	0.3506
10	0.7861	0.2544	0.2427	0.6313	0.3506

Table 3 shows that the second model, with the implementation of ER-Relief algorithm, was able to produce a much higher score compared to Fuzzy C-means. This might be due to the use of ER-Relief as feature selection, which reduces the impact of noise and outliers in the data. It was also seen in the pairwise scatter plot in Figure 10 that there are no noise nor outliers present in the plot, other than some data points that overlap with each other. In this case, the second objective of the study was successfully solved.

Table 4: Clustering Performance Evaluation of ER-PCA-FCM

ER-PCA-FCM					
	Silhouette	Adjusted Rand Index	Homogeneity	Completeness	V-measure
1	0.8166	0.2645	0.2800	0.6125	0.3844
2	0.7817	0.2220	0.2630	0.7906	0.3947
3	0.7823	0.2222	0.2607	0.7818	0.3910
4	0.7667	0.2344	0.2821	0.7097	0.4037
5	0.7654	0.0441	0.0615	0.4281	0.1075
6	0.8002	0.0278	0.0394	0.3630	0.0710
7	0.7654	0.0441	0.0615	0.4281	0.1075
8	0.7817	0.2220	0.2630	0.7906	0.3947
9	0.7654	0.0441	0.0615	0.4281	0.1075
10	0.7548	0.0472	0.0655	0.4384	0.1139

In Table 4, where both ER-Relief and Principal Component Analysis (PCA) were implemented, produced some relatively higher results, however, the consistency was not fulfilled as it has various silhouette scores ranging from 0.75 to 0.80. The ARI, homogeneity, completeness, and V-measure were also considerably low, compared to the previous scores in the previous model. These scores indicate less homogeneity, scattered data points in different clusters, and lower degree of agreement between homogeneity and completeness.

Table 5: Clustering Performance Evaluation of ERPCA-GFCM

ERPCA-GFCM					
	Silhouette	Adjusted Rand Index	Homogeneity	Completeness	V-measure
1	0.7662	0.2591	0.4542	0.5050	0.4782
2	0.7662	0.2591	0.4542	0.5050	0.4782

3	0.7662	0.2591	0.4542	0.5050	0.4782
4	0.7662	0.2591	0.4542	0.5050	0.4782
5	0.7662	0.2591	0.4542	0.5050	0.4782
6	0.7662	0.2591	0.4542	0.5050	0.4782
7	0.7662	0.2591	0.4542	0.5050	0.4782
8	0.7662	0.2591	0.4542	0.5050	0.4782
9	0.7662	0.2591	0.4542	0.5050	0.4782
10	0.7662	0.2591	0.4542	0.5050	0.4782

The enhanced model, which is ERPCA-FCM delivered consistent and good clustering results based on Table 5. The ARI might be a little bit low, but the completeness, v-measure, and homogeneity were moderately good. In conclusion, the enhanced ERPCA-FCM is leading in terms of consistency.

B. Model Performance Evaluation Metrics

To test the overall accuracy of all the0 models, the Random Forest algorithm was used for the training and testing of data. After the dataset was split into training and test set automatically by the said algorithm, a confusion matrix for each model was generated. This confusion matrix was used to solve the Accuracy, Precision, F1 Score, and Recall of each model. These metrics were then used to assess and evaluate the overall performance of each algorithm.

The same approach was implemented in evaluating the model. The model performance of each model after ten (10) tries, to test its consistency and accuracy level.

Table 6: Model Performance Evaluation of FCM

FCM				
	Accuracy	Recall	F1-Score	Precision
1	0.9740	0.9740	0.9738	0.9739
2	0.9740	0.9740	0.9738	0.9739
3	0.9740	0.9740	0.9738	0.9739
4	0.9740	0.9740	0.9738	0.9739
5	0.9740	0.9740	0.9738	0.9739
6	0.9740	0.9740	0.9738	0.9739
7	0.9740	0.9740	0.9738	0.9739
8	0.9740	0.9740	0.9738	0.9739
9	0.9740	0.9740	0.9738	0.9739
10	0.9740	0.9740	0.9738	0.9739

Based on Table 6, the model with the baseline algorithm shows consistent results and has considerably high accuracy results, as well as recall, F1-Score, and Precision. However, it is worth taking note that the clustering results of this model couldn't produce good results, based on Table 2.

Table 7: Model Performance Evaluation of ER-FCM

ER- FCM				
	Accuracy	Recall	F1-Score	Precision
1	0.9750	0.9750	0.9749	0.9749
2	0.9781	0.9781	0.9780	0.9782
3	0.9750	0.9750	0.9749	0.9749
4	0.9781	0.9781	0.9780	0.9782
5	0.9750	0.9750	0.9749	0.9749
6	0.9781	0.9781	0.9780	0.9782
7	0.9781	0.9781	0.9780	0.9782
8	0.9750	0.9750	0.9749	0.9749
9	0.9781	0.9781	0.9780	0.9782
10	0.9750	0.9750	0.9749	0.9749

In Table 7, the second model has somehow inconsistent results and is not as consistent as the first model, which implements the baseline algorithm. The accuracy scores vary from 0.9750 to 0.9871. Recall scores had also the same state. On the other hand, F1 Score has the same results, ranging from two scores which are 0.9780 and 0.9749. Lastly, the precision scores range from 0.9749 to 0.9782.

Table 8: Model Performance Evaluation of ER-PCA-FCM

FCM				
	Accuracy	Recall	F1-Score	Precision
1	0.9750	0.9750	0.9749	0.9749
2	0.9719	0.9719	0.9718	0.9718
3	0.9709	0.9709	0.9707	0.9707
4	0.9719	0.9719	0.9718	0.9718
5	0.9761	0.9761	0.9759	0.9760
6	0.9740	0.9739	0.9738	0.9739
7	0.9740	0.9740	0.9739	0.9739
8	0.9698	0.9698	0.9696	0.9697

9	0.9740	0.9740	0.9739	0.9739
10	0.9761	0.9760	0.9759	0.9760

In Table 8, the scores in all metrics became more inconsistent as the accuracy now ranges from 0.9709, 0.9719, 0.9740, 0.9750, 0.9761, and 0.9698. Of course, this reflects recall, f1-score, and precision as well, producing inconsistent scores. Based on the last 3 tables, it can be seen that as we add more implementation and modifications, the scores became more diverse.

Table 9: Model Performance Evaluation of ERPCA-GFCM

ER- FCM				
	Accuracy	Recall	F1-Score	Precision
1	0.9792	0.9792	0.9790	0.9793
2	0.9792	0.9792	0.9790	0.9793
3	0.9792	0.9792	0.9790	0.9793
4	0.9792	0.9792	0.9790	0.9793
5	0.9792	0.9792	0.9790	0.9793
6	0.9792	0.9792	0.9790	0.9793
7	0.9792	0.9792	0.9790	0.9793
8	0.9792	0.9792	0.9790	0.9793
9	0.9792	0.9792	0.9790	0.9793
10	0.9792	0.9792	0.9790	0.9793

In Table 9, although the clustering produced good moderate results, the ERPCA-GFCM showed consistency with its scores even after numerous executions. The accuracy and precision were also higher compared to all the models. ERPCA-GFCM produced good results and was successful in addressing sensitivity to initial conditions considering that noise and outliers were eliminated, the dimensionality of the date were reduced, there's a good balance between homogeneity and completeness, and the clustering has a degree of success in placing data points in their designated clusters.

The ERPCA-GFCM model demonstrated a strong performance across both model evaluation and clustering evaluation producing an overall accuracy of 98%, surpassing the other models. All of the proposed solutions have successfully addressed their respective problems.

5. CONCLUSION AND RECOMMENDATION

The modifications done in the improved model were able to minimize sensitivity to noise and outliers, with the help of ER-Relief feature selection. FCM's issue in terms of the dimensionality of data was also addressed successfully using

Principal Component Analysis (PCA). In addition, the application of Global Fuzzy C-means was able to produce good and consistent results and was able to solve the problem of the baseline algorithm in dealing with initial conditions.

The first solution, which involved implementing the GFCM Algorithm, effectively addressed the intended problem, which is FCM's sensitivity to initial conditions, based on its consistent results and high accuracy. However, it is worth noting that this approach appears to be more time-consuming compared to the original algorithm. In this case, running time is one of the challenges we were able to identify during the process. Apart from this, the ER-relief algorithms were also able to solve the problem with noise and outliers. This was proven by how it selected features based on their relevance, which lessened the impact of noises as it eliminated the rest of the features. The outliers were also omitted, and the application of principal component analysis was able to reduce the dimensionality of the data from initial number of 88 to 5.

Although there are identified issues in terms of running time and low clustering results, the upgraded model performed relatively well overall and had high consistency and accuracy results, making it a good model for future researchers to use and improve.

The ERPCA-GFCM model demonstrated a strong performance across both model evaluation and clustering evaluation producing an overall accuracy of 98%, surpassing the other models. All of the proposed solutions have successfully addressed their respective problems.

There are still plenty of opportunities for the model's improvement, particularly in enhancing the model's clustering performance. To further assess the model's performance, future researchers can apply the enhanced model on different datasets to test how it can handle much bigger data that has high dimensionality and broader features.

High clustering results can also greatly affect the execution of ERPCA-GFCM. Even though the results from this study were considered moderate, it is still highly recommended by the researchers to implement other parameters such as initial and comprehensive data cleaning and pre-processing, removing null values, iterative refinement, and calculating weights of the features from the dataset in order to improve the relevance sorting. Further visualization can also be implemented by using self-organized maps (SOM) to identify and analyze data dimensions.

In addition, solving the problem of the enhanced model in regard to the running time can be solved using the Fast Global Fuzzy C-means. FGFCM is another variant of enhanced Fuzzy C-means that has the main purpose of improving the convergence rate or speed of Global Fuzzy C-means without affecting its initial results and quality (Li et al., 2006).

By addressing these areas of improvement, it is possible to enhance the overall effectiveness and accuracy of the model's clustering functionality, leading to more reliable and insightful analyses.

ACKNOWLEDGEMENT

To all the people who contributed to the success of this study, your contributions are highly appreciated. The researchers are deeply grateful for your presence in our lives and for being an integral part of this journey. Thank you for your trust and unwavering support.

6. REFERENCES

[1] Aid., "Internet of things (IOT) security: Challenges and best practices," Apriorit, Dec. 2022.

[2] Al-Zoubi, M.D.B., A.D. Ali and A.A. Yahya, "Fuzzy clustering-based approach for outlier detection," Proceedings of the 9th WSEAS International Conference on Applications of Computer Engineering, pp. 192-197, March 2010.

[3] Askari, S., "Fuzzy C-Means clustering algorithm for data with unequal cluster sizes and contaminated with noise and outliers: Review and development," Expert Systems with Applications, vol. 165, March 2021.

[4] Archana, T., Sachin, D., "Dimensionality Reduction and Classification through PCA and LDA," International Journal of Computer Applications, 0975-8887, vol. 122, no. 17, July 2015.

[5] Bezdek, J., Ehlrich, R., Full, W., "FCM: The fuzzy c-means clustering algorithm," Computers & Geosciences, vol. 10, pp. 191-203, 1984.

[6] Bobrovnikova, K., Lysenko, S., & Savenko, O., "DDoS Botnet Detection Technique Based on the Use of the Semi-Supervised Fuzzy C-Means Clustering," 14th International Conference on ICT in Education, Research

and Industrial Applications. Integration, Harmonization and Knowledge Transfer, vol. 2: Workshops, vol. 2104, pp. 688-695, 2018.

[7] Chen, Y., Hou, J., Li, Q., & Long, H., “DDoS Detection Based on Random Forest,” IEEE International Conference on Progress in Informatics and Computing (PIC), 2020.

[8] C. Kolias, G. Kambourakis, A. Stavrou and J. Voas, "DDoS in the IoT: Mirai and Other Botnets," in Computer, vol. 50, no. 7, pp. 80-84, 2017.

[9] C. Y. Lin, M. Wu, J. A. Bloom, I. J. Cox, and M. Miller, “Rotation, scale, and translation resilient public watermarking for images,” IEEE Trans. Image Process., vol. 10, no. 5, pp. 767-782, May 2001.

[10] “DDoS attacks in 2022: Trends and obstacles amid worldwide political crisis,” Infosecurity Magazine, https://www.infosecurity-magazine.com/blogs/ddos-attacks-in-2022-trends/ (accessed Jun. 2023).

[11] “DDoS Year-in-Review Report by StormWall,” https://stormwall.network/ddos-report-stormwall-2022, 2022.

[12] S. Deng, “Clustering with fuzzy C-means and common challenges,” Journal of Physics: Conference Series, vol. 1453, no. 1, p. 012137, 2020.

[13] Devi, M., Priya, V., Sarumathy, S., & Sujatha, R., ” Preventing DDoS Attack Using Fuzzy C Mean Clustering. International Journal of Innovative Research in Science, Engineering and Technology, 2016.

[14] Gupta, A., “Fuzzy C-means clustering (FCM) algorithm in Machine Learning,” Jan. 2023.

[15] I. Sharafaldin, A. H. Lashkari, S. Hakak and A. A. Ghorbani, "Developing Realistic Distributed Denial of Service (DDoS) Attack Dataset and Taxonomy," 2019 International Carnahan Conference on Security Technology (ICCST), Chennai, India, pp. 1-8, 2019.

[16] Jing, H., & Wang, J., “Detection of ddos attack within industrial IOT devices based on clustering and graph structure features,” Security and Communication Networks, 2022.

[17] Lee, S.-H., Shiue, Y.-L., Cheng, C.-H., Li, Y.-H., & Huang, Y.-F., “Detection and Prevention of DDoS Attacks on the IoT,” Applied Sciences, 12(23), 12407, 2022.

[18] Lunkad, D., & Singh, G., “DDOS Attack Detection Using Machine Learning For Network Performance Improvement,” International Journal of Creative Research Thoughts (IJRCT), 8(9), 1782–1786, 2020.

[19] Mittal, M., Kumar, K., & Behal, S., “Deep learning approaches for detecting ddos attacks: A systematic review,” Soft Computing, 2022.

[20] Munandar, Tb & Musdholifah, A., “Comparative Study Between Primitive Operation Complexity Against Running Time Application On Clustering Algorithm,” International Journal of Advanced Research in Computer Science. vol. 5, pp. 164 – 169, 2014.

[21] Najafimehr, M., Zarifzadeh, S., & Mostafavi, S., “A hybrid machine learning approach for detecting unprecedented ddos attacks,” The Journal of Supercomputing, vol. 78(6), pp. 8106–8136, 2022.

[22] Qian, J., Nguyen, N. P., Oya, Y., Kikugawa, G., Okabe, T., Huang, Y., & Ohuchi, F. S., “Introducing self-organized maps (SOM) as a visualization tool for Materials Research and Education,” Results in Materials, 4, 100020.

[23] Khantimirov, R., “DDoS activity intensifies in Asia Pacific,” Infosecurity Magazine, https://www.infosecurity-magazine.com/blogs/ddos-activity-intensifies-in-asia/ (accessed Jan. 2023)

[24] Rathore, S., & Park, J. H., “Semi-supervised Learning Based Distributed Attack Detection Framework for IOT,” Applied Soft Computing, vol. 72, pp. 79–89, 2018.

[25] Seifousadati, A., Ghasemshirazi, S., & Fathian, M., “A Machine Learning Approach for DDoS Detection on IoT Devices,” 2021.

[26] Shen, Y., E, H., Chen, T., Xiao, Z., Liu, B., & Chen, Y., “ High-dimensional data clustering with fuzzy C-means: Problem, reason, and solution,” Advances in Computational Intelligence, pp. 89–100, 2021.

[27] Siringoringo, R., & Jamaluddin, J., “Initializing the fuzzy C-means cluster center with particle swarm optimization for sentiment clustering,” Journal of Physics: Conference Series, 1361(1), 012002, 2019.

[28] Suresh, M., & Anitha, R., “Evaluating machine learning algorithms for detecting ddos attacks,” Advances in Network Security and Applications, pp. 441-452, 2011.

[29] Wang, W., Zhang, Y., Li, Y., & Zhang, X., “The global fuzzy C-means clustering algorithm. 2006 6th World Congress on Intelligent Control and Automation,” 2006.

[30] Wiharto, W., & Suryani, E., “The comparison of clustering algorithms K-means and fuzzy c-means for segmentation retinal blood vessels,” Acta Informatica Medica, vol. 28(1), pp. 42, 2020.

[31] Winkler, R., Klawonn, F., & Kruse, R., “Fuzzy c-means in high dimensional spaces,” International Journal of

Fuzzy System Applications, vol.1(1), pp. 1–16, 2011.

[32] Winkler, R., Klawonn, F., & Kruse, R., “Problems of fuzzy c-means clustering and similar algorithms with high dimensional data sets,” Challenges at the Interface of Data Analysis, Computer Science, and Optimization, pp. 79–87, 2012.

[33] Yang, SH., Hu, BG., “Efficient Feature Selection in the Presence of Outliers and Noises,” Information Retrieval Technology. AIRS 2008. Lecture Notes in Computer Science, vol. 4993.

DOI: https://doi.org/10.15379/ijmst.v10i2.3298

This is an open access article licensed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0/), which permits unrestricted, non-commercial use, distribution and reproduction in any medium, provided the work is properly cited.