ISSN: 2581-3404 (Online) IF: 5.68 (SJIF)
IJIRTM, Volume-9, Issue-3, June-2025

Privacy Preserving Clustering in Data Mining Using Piecewise Vector Quantization Approach

[1]Shikha Jawre, [2]Pradeep Pandey
[1]Research Scholar, [2]Assistant Professor
[1]Department of Computer Science & Engineering,
[1] SAM College of Engineering and Technology, Bhopal, India.

Abstract. *A large volume of detailed personal data—such as shopping habits, criminal records, medical histories, and credit information—is routinely collected and shared for various data mining applications. On one hand, this data represents a valuable asset for businesses and government agencies, enabling informed decision-making through comprehensive analysis. On the other hand, privacy regulations and growing concerns over individual data protection can hinder data sharing and limit its use in analytics. To address the challenge of balancing data utility with privacy preservation, this work proposes a vector quantization-based approach for privacy-preserving data clustering. The method involves segmenting each row of the dataset into multiple parts (piecewise segmentation), followed by applying K-Means quantization to each segment individually. The quantized segments are then recombined to form a transformed dataset suitable for clustering while maintaining privacy. Experimental results are presented to evaluate the performance of the proposed method. These experiments aim to identify the optimal segment size and quantization parameter that strike the best balance between clustering accuracy and data privacy. The results demonstrate that appropriate tuning of these parameters can significantly improve the trade-off, enabling effective data analysis without compromising sensitive information.*

Keywords: Data Mining (DM), Knowledge, classification, Learning Analytics (LA), Water Treatment database and F measure.

Introduction

Over the past two decades, the volume of personal data collected about individuals has grown significantly. This data originates from diverse sources such as medical records, financial transactions, library usage, phone logs, and shopping behaviours. Advances in database systems, networking infrastructure, and computational power have enabled the integration and digital analysis of this vast information. While this development has fuelled the rise of data mining tools capable of extracting valuable insights and trends, it also raises serious concerns regarding individual privacy. The ease of accessing and analysing such detailed personal data can lead to misuse or unintended disclosure, creating a conflict between the utility of data mining and the need to protect sensitive information. In this paper, we address this challenge by introducing a piecewise vector quantization approach for privacy-preserving clustering. The proposed method segments

each data instance into smaller parts and applies quantization to each segment independently using the K-Means algorithm. These transformed segments are then recombined to form a modified dataset suitable for clustering, while reducing the risk of exposing sensitive information. The effectiveness of this approach is evaluated by analyzing the trade-off between clustering accuracy and data privacy, aiming to achieve a balance that supports both data utility and confidentiality.

Data mining

Data mining is a powerful technique focused on extracting hidden, predictive patterns from large datasets. It employs advanced algorithms to analyze vast volumes of data and identify relevant and useful information. These tools enable the prediction of future trends and behaviors, empowering organizations to make informed, proactive, and knowledge-driven decisions. With the volume of data doubling annually, the ability to effectively mine this data has become increasingly essential for transforming raw data into actionable insights. The development of data mining has evolved over time, beginning with the storage of business data on computers, advancing through improvements in data access, and culminating in the creation of technologies that allow users to explore data in real time. Unlike traditional data access methods that are retrospective in nature, data mining offers prospective analysis—enabling not just understanding of what has happened, but also what is likely to happen next. The widespread application of data mining in the business sector has been made possible by the convergence of three mature technologies: massive data collection and storage infrastructure, powerful computational capabilities, and sophisticated data analysis algorithms. Together, these advancements have positioned data mining as an essential tool for strategic decision-making in the modern data-driven world.

- Massive data collection
- Powerful multiprocessor computers
- Data mining algorithms

Scope of data mining

The term data mining draws an analogy to the process of extracting valuable minerals from mountains—just as miners sift through vast quantities of earth to find precious ore, data mining involves exploring large databases to uncover useful business insights. Whether through exhaustive scanning or intelligent algorithms, the goal is to locate high-value information hidden within massive datasets. When applied to sufficiently large and high-quality databases, data mining technologies offer significant business advantages. One key capability of data mining is the automated prediction of trends and behaviours. Traditionally, answering such questions required extensive manual analysis; now, data mining enables rapid and automated responses directly from the data. A common application is targeted marketing, where past campaign data is used to identify customers most likely to respond positively to future promotions, thereby maximizing return on investment. Other predictive applications include forecasting bankruptcy, detecting credit default, and identifying customer segments with similar behavioural patterns. Another crucial function is the automated discovery of previously unknown patterns. Data mining tools can analyse complex datasets to reveal hidden relationships in a single step. For example, analysing retail sales data might uncover that seemingly unrelated products are frequently purchased together. Additional applications include detecting fraudulent transactions, flagging anomalies, and identifying data entry errors.

Applications of data mining

There is a rapidly growing body of successful applications in a wide range of areas as diverse as: analysis of organic compounds, automatic abstracting, credit card fraud detection, financial forecasting, medical diagnosis etc. Some examples of applications (potential or actual) are:

- A supermarket chain mines its customer transactions data to optimize targeting of high value customers
- A credit card company can use its data warehouse of customer transactions for fraud detection
- A major hotel chain can use survey databases to identify attributes of a 'high-value' prospect.

Literature Survey

Numerous studies have explored the application of data mining techniques—particularly clustering and classification—for improving academic performance, some of which provide foundational insights for the implementation of the Piecewise Vector Quantization (PVQ) approach.

Md. Hedayetul Islam Shovon and Mahfuza Haque (2018) proposed a method to enhance student academic performance by applying K-means clustering and decision tree algorithms. Their study applied data mining techniques on student databases to predict learning behaviors, ultimately aiming to reduce failure rates and support educators in timely intervention.

Dr. Priyanka Sharma (2017) focused on distributed data mining for performance prediction. Her work demonstrated that distributing the training and testing phases across different nodes can enhance classification accuracy for large datasets. This approach supports the early identification of at-risk students, contributing positively to both individual outcomes and institutional performance.

M.I. López et al. (2012) introduced a classification-via-clustering approach to predict student grades based on forum participation. The study used the EM clustering algorithm and highlighted the potential of automating forum message evaluation through text mining. Future work includes building network analysis tools for visualizing forum interactions.

Mahesh Singh, Anita Rani, and Ritu Sharma (2014) applied K-means clustering using the Weka interface to evaluate student learning activities. Their results suggest that clustering helps identify learning patterns, enabling timely instructional interventions to enhance educational quality.

Brijesh Kumar Baradwaj and Saurabh Pal (2011) employed decision tree classification to analyze student performance based on parameters like attendance, class tests, seminars, and assignments. The model aimed to predict end-semester divisions and identify students requiring additional academic support. These studies collectively emphasize the potential of clustering and classification models in educational settings. They provide a strong basis for implementing advanced methods like PVQ to achieve improved accuracy in predictive analytics while addressing privacy preservation and data sensitivity challenges.

Alhayan et al. (2025) propose an advanced approach to anomaly-based network intrusion detection by integrating an improved Snow Ablation Optimizer (ISAO) with dimensionality reduction and a hybrid deep learning model. The study focuses on addressing challenges posed by high-dimensional network data and complex attack patterns, which often reduce the effectiveness and efficiency of intrusion detection systems (IDS). The ISAO algorithm is employed to optimize feature selection, effectively reducing data dimensionality while retaining critical information for detecting anomalies. This step enhances computational efficiency and detection accuracy by eliminating redundant and irrelevant

features. The hybrid deep learning model combines the strengths of different architectures, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), to capture both spatial and temporal patterns in network traffic. This combination enables the system to identify sophisticated and previously unknown attack vectors. Experimental results on benchmark datasets reveal that the proposed method achieves superior detection rates and reduced false alarms compared to existing techniques. The study highlights the potential of combining meta-heuristic optimization with deep learning to develop robust, scalable, and adaptive IDS for modern network environments.

Problem Statement

The goal of privacy-preserving clustering is to protect the underlying attribute values of objects subjected to clustering analysis. In doing so, the privacy of individuals would be protected. The problem of privacy preservation in clustering can be stated as follows: Let D be a relational database and C a set of clusters generated from D. The goal is to transform D into D' so that the following restrictions hold:

- A transformation T when applied to D must preserve the privacy of individual records, so that the released database D' conceals the values of confidential attributes, such as salary, disease diagnosis, credit rating, and others.
- The similarity between objects in D must be the same as that one in D', or just slightly altered by the transformation process. Although the transformed database D' looks very different from D, the clusters in D and D' should be as close as possible since the distances between objects are preserved or marginally changed.

Our work is based on piecewise Vector Quantization method and is used as non-dimension reduction method. It is modified form of piecewise vector quantization approximation which is used as dimension reduction technique for efficient time series analysis.

Methodology

Series of experiment was performed varying segment size(L) i.e. segment number (w) varies and varying number of cluster for quantization(K). Our evaluation approach focused on the overall quality of generated clusters after transforming dataset and the distortion produced in the dataset. Experiment was based on following steps

- We modified the dataset by dealing with the missing value. To do so we replace it with average value of that attribute over the whole dataset.
- We applied piecewise vector quantization method to transform dataset.
- We selected K means to find the clusters in our performance evaluation. Our selection was influenced by following aspects (a) K-means is one of the best known clustering algorithm and is scalable. (b) K-means was also used in our codebook generation step. Number of cluster to be find from original and transformed dataset was taken same as number of cluster for quantization. This is the limitation in our experiment, experiment can also be used to find result using two different value, one for number of cluster for quantization(K) and other for number of cluster to be find from original and transformed dataset. Although we performed experiment taking same value.
- We compared how closely each cluster in the transformed dataset matches its

corresponding cluster in the original dataset. We expressed the quality of the generated clusters by computing the F-measure.

- We compared the distortion produced due to transformation of dataset by the distortion metric.

Experiment Analysis

Dataset Information

Water treatment dataset available on UCI Machine Learning Repository was taken for experimenting. It consists of 527 records and 38 attributes. Attributes type is integer or real.

Table 1: Dataset Information.

Feature	**Description**
Dataset Name	**Water Treatment Plant Dataset**
Source	**UCI Machine Learning Repository**
Total Number of Instances (Records)	**527**
Number of Attributes (Features)	**38**
Attribute Type	**Real and Integer (Continuous Numeric)**
Missing Values	**Present in several attributes**
Application Domain	**Environmental Monitoring / Industrial Control**
Dataset	**Water Treatment**
No .of records	527
No. of attributes	38

Steps for Dataset Procedure

Series of experiment was performed varying segment size (L) i.e. segment number (w) varies and varying number of cluster for quantization (K). Our evaluation approach focused on the overall quality of generated clusters after transforming dataset and the distortion produced in the dataset. Experiment was based on following steps

1. We modified the dataset by dealing with the missing value. To do so we replace it with average value of that attribute over the whole dataset.

2. We applied piecewise vector quantization method to transform dataset.

3. We selected K means to find the clusters in our performance evaluation. Our selection was influenced by following aspects (a) K-means is one of the best known clustering algorithm and is scalable. (b) K-means was also used in our codebook generation step. Number of cluster to be find from original and transformed dataset was taken same as number of cluster for quantization. This is the limitation in our experiment, experiment can also be used to find result using two different value, one for number of cluster for quantization (K) and other for number of cluster to be find from original and transformed dataset. Although we performed experiment taking same value.

4. We compared how closely each cluster in the transformed dataset matches its corresponding cluster in the original dataset. We expressed the quality of the generated clusters by computing the F-measure.

5. We compared the distortion produced due to transformation of dataset by the distortion metric.

Results

Experiment was performed for measuring distortion in transformed dataset on different K value keeping the L constant and on different L value keeping the K constant. The result which came from our experiment is shown in Table 2 and corresponding graph between distortion and K and between distortion and L value is drawn as shown next pages.

Table 2: Distortion value at different K and L value.

L/K	5	10	15	20	25	30	35	40	45	50
2	42.810	22.840	14.070	11.800	11.040	11.050	10.240	9.023	8.490	7.949
3	42.830	22.880	14.170	11.880	11.150	11.110	10.330	9.140	8.610	8.050
4	43.010	23.162	14.860	12.300	11.560	11.570	10.820	9.670	9.130	8.420
5	43.350	23.450	14.960	12.780	12.240	11.980	11.420	10.250	9.790	9.330
6	43.240	23.550	15.200	13.390	12.750	12.470	11.580	10.360	10.050	9.430
7	43.250	23.590	15.230	13.410	12.770	12.520	11.610	10.350	10.070	9.730
8	43.270	23.710	15.330	13.560	12.930	12.620	11.750	10.420	10.230	9.880
9	44.550	25.520	17.930	16.170	15.360	13.310	14.320	12.500	13.110	12.150
10	44.600	25.600	18.040	16.270	15.440	15.380	14.420	12.890	13.160	12.210
11	44.540	25.670	18.100	16.450	15.600	15.490	14.470	13.020	12.590	12.560
12	44.547	25.720	18.190	16.580	15.740	15.650	14.410	13.370	13.460	12.670
13	44.548	25.690	18.200	16.587	15.700	15.360	14.420	13.440	13.420	12.580
14	44.548	25.695	18.210	16.580	15.690	15.380	14.420	13.390	13.440	12.582
15	45.620	27.210	21.320	18.030	17.330	16.790	15.920	14.554	14.130	14.012
16	45.625	27.211	21.327	18.036	17.350	16.800	15.920	14.557	14.590	14.020
17	45.628	27.217	21.330	18.033	17.300	16.760	16.000	14.270	14.620	14.040
18	45.638	27.288	21.340	18.480	17.330	16.960	16.110	14.680	14.750	14.100
19	45.639	27.232	21.350	18.392	17.190	16.910	16.200	14.690	14.760	14.150

Figure 1 shows Distortion Vs Segment Size (L). Segment size varies from 2 to 19 and its corresponding distortion measured by distortion metric on various values of K is shown. It can be easily concluded that distortion increases with increase in L and it's obvious as more the value of L more the attribute is affecting for quantization so more is the irregularity and more the distortion.

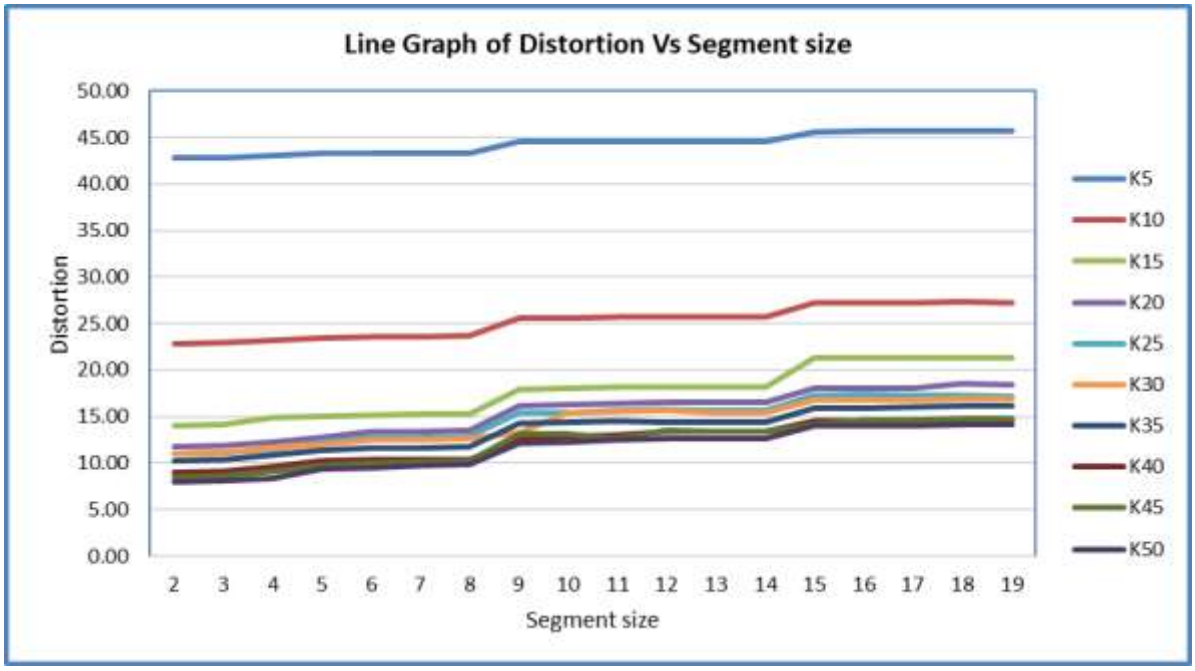

Figure 1: Line Graph of Distortion Vs Segment size(L) at different K value.

Distortion also reduces with increase in K as with increase in K less number of row points are used for quantization (as average numbers of points per cluster reduces and codebook generation that is later used for quantization, take place as mean of all points falling in a cluster) so less is the distortion. It's also shown in Figure 2 as a line graph and Figure 3 as bar graph between Distortion and number of cluster for quantization (K) at various segment size values, Distortion leads to loss of information which can leads to loss in information in cluster. So it should be reduced.

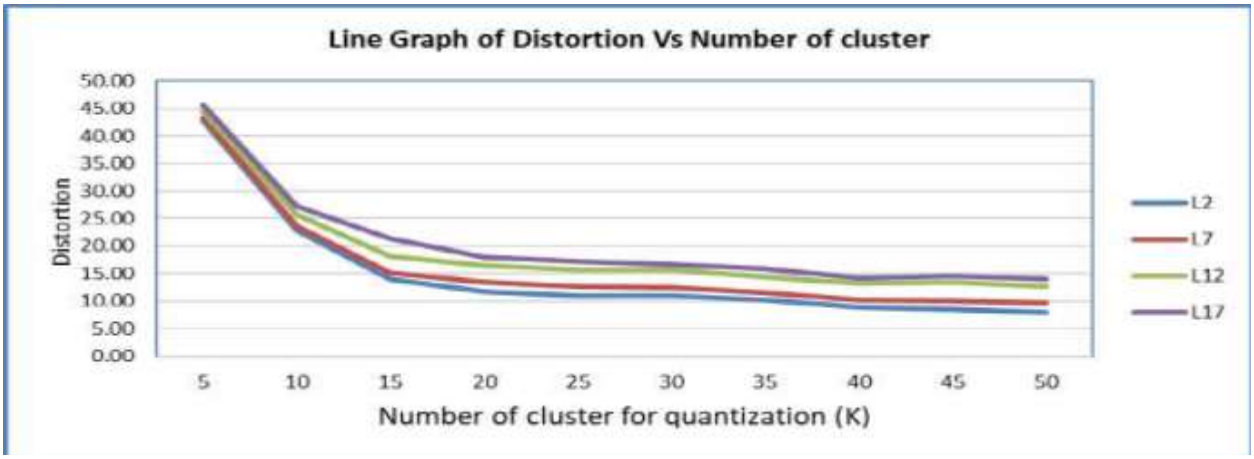

Figure 2: Line Graph of Distortion Vs Number of cluster for quantization(K) at different L.

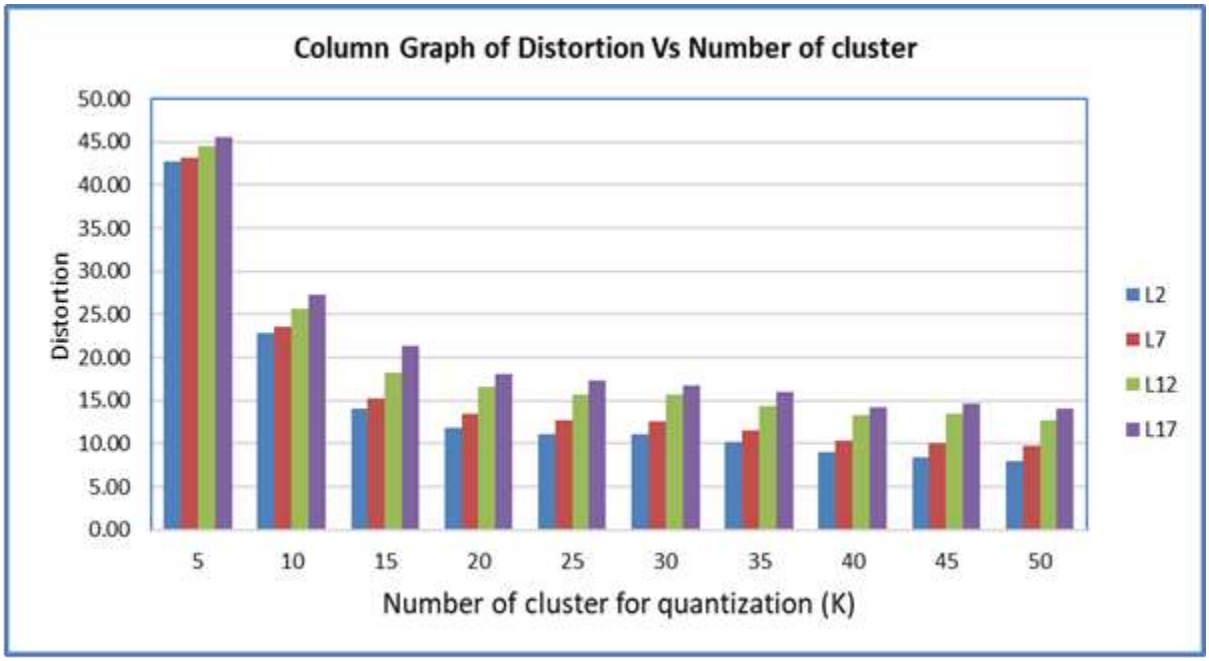

Figure 3: Column Graph of Distortion Vs Number of cluster for quantization(K) at different L.

Conclusion

We have showed analytically and experimentally that Privacy-Preserving Clustering is to some extent possible using piecewise vector quantization approach. To support our claim, we used water treatment dataset available on UCI Machine Learning Repository and performed experiment on it varying segment size and number of cluster for quantization. We evaluated our method taking into account two important issues: distortion and Fmeas Our experiment showed the variation of Fmeasure and distortion with segment size and number of cluster for quantization. It was found optimum segment size and number of cluster for quantization which give nice Fmeasure and distortion and in turn privacy for water treatment dataset

The proposed method was experimentally evaluated on the Water Treatment dataset using two primary metrics:

- F-measure, to assess the quality of clustering post-transformation.
- Distortion, to quantify the extent of data transformation and thus the level of privacy.

Key findings from the experiments include:

- Distortion increases with segment size (L) and decreases with a higher number of clusters (K), confirming the balance between privacy and reconstruction accuracy.
- F-measure improves with increasing K and reaches optimal performance at K = 30 and L = 9, with F-measure ≈ 0.780 and Distortion ≈ 13.31.

These results demonstrate that the proposed PVQ method effectively balances the trade-off between privacy preservation and clustering performance, making it a viable solution for secure data mining in privacy-sensitive domains.

References

1. P. Ferragina, “String algorithms and data structures,” arXiv (Cornell University), Jan. 2008, doi: 10.48550/arxiv.0801.2378.
2. A. Anchlia, “Enhancing Query Performance Through Relational Database Indexing,” International Journal of Computer Trends and Technology, vol. 72, no. 8, p. 130, Aug. 2024, doi: 10.14445/22312803/ijctt-v72i8p119.
3. Q. M. Alzubi, M. Anbar, Z. N. M. Alqattan, M. A. Al-Betar, and R. Abdullah, “Intrusion detection system based on a modified binary grey wolf optimisation,” Neural Computing and Applications, vol. 32, no. 10, p. 6125, Feb. 2019, doi: 10.1007/s00521-019-04103-1.
4. P. Beaumont and M. Huth, “Constrained Bayesian Networks: Theory, Optimization, and Applications,” arXiv (Cornell University), Jan. 2017, doi: 10.48550/arxiv.1705.05326.
5. W. Ziarko, “Variable precision rough set model,” Journal of Computer and System Sciences, vol. 46, no. 1, p. 39, Feb. 1993, doi: 10.1016/0022-0000(93)90048-2.
6. A. Mukherjee and A. K. Das, “Einstein-operations on fuzzy soft multi sets and decision making,” Boletim da Sociedade Paranaense de Matemática, vol. 40, p. 1, Feb. 2022, doi: 10.5269/bspm.32546.
7. K. A. Dhanya, S. Vajipayajula, K. Srinivasan, A. Tibrewal, S. K. Thangavel, and T. G. Kumar, “Detection of Network Attacks using Machine Learning and Deep Learning Models,” Procedia Computer Science, vol. 218, p. 57, Jan. 2023, doi: 10.1016/j.procs.2022.12.401.
8. I. S. Thaseen, B. Poorva, and P. S. Ushasree, “Network Intrusion Detection using Machine Learning Techniques,” 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), p. 1, Feb. 2020, doi: 10.1109/ic-etite47903.2020.148.
9. T. Khorram, “Network Intrusion Detection using Optimized Machine Learning Algorithms,” European Journal of Science and Technology, Jun. 2021, doi: 10.31590/ejosat.849723.
10. A. M. Bamhdi, I. Abrar, and F. Masoodi, “An ensemble based approach for effective intrusion detection using majority voting,” TELKOMNIKA (Telecommunication Computing Electronics and Control), vol. 19, no. 2, p. 664, Feb. 2021, doi: 10.12928/telkomnika.v19i2.18325.
11. P. M. Corea, Y. Liu, J. Wang, S. Niu, and H. Song, “Explainable AI for Comparative Analysis of Intrusion Detection Models,” arXiv (Cornell University), Jun. 2024, doi: 10.48550/arxiv.2406.09684.
12. R Dubey, “An empirical study of intrusion detection system using feature reduction based on evolutionary algorithms and swarm intelligence methods”, International Journal of Applied Engineering Research, 2017. pp. 8884-8889.
13. P. Dini, A. Elhanashi, A. Begni, S. Saponara, Q. Zheng, and K. Gasmi, “Overview on Intrusion Detection Systems Design Exploiting Machine Learning for Networking Cybersecurity,” Applied Sciences, vol. 13, no. 13, p. 7507, Jun. 2023, doi: 10.3390/app13137507.
14. D Rathore, Praveen Mannepalli, “Diseases Prediction and Classification Using Machine Learning Techniques”, AIP Conference Proceedings, 2022, pp. 1-8.
15. W. H. Aljuaid and S. S. Alshamrani, “A Deep Learning Approach for Intrusion Detection Systems in Cloud Computing Environments,” Applied Sciences, vol. 14, no. 13, p. 5381, Jun. 2024, doi: 10.3390/app14135381.