Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 15, Issue. 2, No.3 : 2024
ISSN : **1906-9685**

EXCLUSIVE ADAPTIVE WEIGHTED SUPPORT VECTOR MACHINE CLASSIFICATION METHOD FOR PRIVACY PRESERVING IN CLOUD OVER BIG DATA USING HADOOP FRAMEWORK

Dr.A.Kanimozhi, Assistant Professor, Department of Computer Science, Sri Krishna Adithya College of Arts and Science, Coimbatore.
S. Suhaibu, Department of Computer Science, Sri Krishna Adithya College of Arts and Science, Coimbatore.

Abstract:
One of the most important aspects of big data research is data security. Applications for the cloud system handle private data, including medical, business, and personal records. The cloud systems that contain this kind of data may be at risk from threats. However, large data migration cannot be protected by conventional security methods. Large volumes of data are created, and the security aspects of that data are handled throughout the cloud by an efficient privacy-preserving system. The Hadoop framework is used to first cluster and balance the data in a cloud-based dataset. Exclusive Enhanced Word Auto Key Encryption (WAKE) and Adaptive Weighted Support Vector Machine with Continuous scatter search (CSS) Optimization Algorithm classifiers, known as AWSVM-CSS technique, respectively, encrypt and classify the clustering process, which is carried out by the Density Peak Weighted Fuzzy C-means Clustering (DPWFCM) algorithm and processed by the Hadoop framework. A hybrid technique is used to carry out the encryption and decryption processes. According to experimental observations, the suggested method is very successful in terms of classification and encryption.

Keywords Big data · Cloud computing · Data security · Encryption · Privacy · Decryption · Clustering and classification

1. Introduction

The technology known as cloud computing enables users to easily store and process data. Service delivery, development processes, and industrial infrastructure were all altered by cloud computing. Predictive models, data analytics, online banking, and basic data storage are just a few of the many functional areas of cloud computing. Cloud computing comes in three varieties: public, private, and hybrid. Private cloud data is more secure with certain safeguards, but public cloud data is accessible to anyone who might be concerned about privacy [2]. The combined use of private and public clouds is known as hybrid cloud [14]. Numerous problems and difficulties pertaining to the security model may arise when data is transmitted via cloud architecture. This same security concern in the hybrid cloud environment was also addressed in the proposed study. Instead of constructing and maintaining computing infrastructure in-house, cloud computing enables the efficient use of resources through a virtual machine, storage, or services dependent on customer demand [10].

Users' data is kept on external storage managed by cloud service providers and accessible online in a hybrid cloud. Additionally, this private information may be stored by one supplier or by several providers. Only those who are permitted can access data stored in the cloud thanks to confidentiality. Unreliable cloud service providers' (CSPs') methods can compromise confidentiality. Confidentiality is therefore a major security concern for hybrid cloud computing [17].

When multiple substances need to exchange information, there needs to be a way to limit who can access that information. Many methods are being proposed to protect the content of information and prevent unauthorized individuals from accessing and discovering it by using access control, while permitting many authorized individuals to share the information [23]. SaaS will be used to collect and store security data from the company's processes. Data must be protected online to prevent private information from leaking. To keep the network safe for traffic, there should be a robust encryption system in place. Human mistake is responsible for the majority of the information in the Association. There are bad people inside, and they can do a lot of damage [25]. A malicious internal person could be a contract employee, a former or present employer, or a partner company that obtains information and purposefully uses, steals, or violates it [13]. In order to overcome resistance, safe techniques are used, such as gaining firm control and constantly experimenting with the functionality of external forms and screens [6, 9]. The combination of two or more encryption algorithms is known as hybrid file encryption. It leverages the strengths of all encryption types by combining symmetric and asymmetric encryption [20].

1.1 Motivation

A crucial component of the cloud services offered globally is security. In a cloud network environment, data security is crucial (https://www.unb.ca/cic/datas ets/nsl.html). Security risks include things like application encryption, security, security key management, identification, access management, audit schedules, and physical and user access control [8]. In a hybrid cloud, two of the most important security concerns are data encryption and user access control. Security issues can be lessened by protecting cloud data and managing data access. The business model for cloud-based services is growing from basic data storage to essential internet banking, data analytics, and predictive modeling.

Consequently, a thorough examination of various data security models is carried out in a cloud setting in order to suggest an effective data security strategy to accomplish the intended advantages of network scalability and data security [22]. The categorization process has become necessary due to the identification of various types of network infiltration, and the volume of data necessitates the use of Hadoop mechanisms for data management.

1.1 Aim and objectives

The purpose and goals of the suggested methods are

• The input dataset is initialized, and the Density Peak Weighted Fuzzy C-means Clustering (DPWFCM) algorithm is used to cluster the data.

• Map reduce programming representation is typically used in Hadoop frameworks to extend highly comparable functions that advance and generate large amounts of data.

• Exclusive Enhanced Word Auto Key Encryption (WAKE) is used to perform the encryption process.

• The Adaptive Weighted Support Vector Machine with Continuous Scatter Search (CSS) Optimization Algorithm classifier, known as AWSVM-CSS, completes the security evaluation process through encryption. In the event that the encrypted data is not correctly encrypted, the convolution process is applied once more.

The article's remaining content is arranged as follows: Section 2 presents the current cloud and big data-based security schemes; Section 3 presents the suggested clustering and classification scheme; Section 4 illustrates the results of DPWFCM, Exclusive Enhanced WAKE, and AWSVM-CSS with Hadoop framework; and Section 5 concludes the article with future work.

2 Related works

Mushtaqet al. (2017) [26] introduced quad layered outline for securing the information, data confidentiality, data breaches, and process linked aspects. Through this layered design, they have preserved the confidentiality of private data and attempted to build the trust of the user in cloud computing [19]. This layered outline averted the private data by way of several means, i.e., Secure

Data Transmission, Encrypted Data, and it ‘s Handling, etc.

In a fully functional engine, the Advance Encryption Standard (AES) is used to encrypt and decrypt data. In both encryption and decryption processes, it has the capacity to deliver data at a high pace [4]. The data is secured using the Triple Data Encryption Standard (DES), which ciphers the information three times [1]. RC2 is a type of encryption system in which data is encrypted and decrypted [12].

The Hadoop security framework is used to secure data in the big data environment [5]. These systems have a number of flaws, including the installation of appropriate security mechanisms over large amounts of data, as well as the clustering and classification pro- cesses [11].

As the popularity of the internet grows among users throughout the globe, so the need of maintaining security and keeping the system informed of dangerous activity [21]. It has the following shortcomings, as stated by the classic detection scheme: high
false alarm rate, low detection rate, and so on. The classifier’s performance is deter- mined by the required of its terms of usefulness [24].

For the progress of business, the computing sector has already begun to use cloud infrastructure. Cloud-based infrastructure and apps are already transforming businesses across all industries and large corporations [18]. The data generated by these cloud- based infrastructure and apps is enormous. This massive volume of data can be used to forecast the future [15]. From the literature it is identified that the SVM is effective and parameter setting is deployed to fine tune the process.

3 Proposed methodology

The Fuzzy C-Means (FCM) technique is sensitive to particular cluster centers, the cluster centers and cluster count must be chosen at random. Common problems with the method include low stability, delayed convergence, local best solution, and several clustering attempts. Consequently, the density peak-based weighted FCM algorithm, or DPWFCM algorithm, is created. The density distance x is indicated by two factors, px and ðx, which initially establish the cluster centers.

$$_x = p_x \hat{d}_x$$

Determine the x values of each point in the cluster by traversing the k sample points. The first p points are obtained by sorting the density distance values in descending order. The following formula is used to estimate an average density distance:

By iteratively optimizing the objective function, the weighted FCM algorithm, an unsupervised learning method based on objective function optimization, divides the KDD dataset. The clustering problem is examined via the FCM technique using fuzzy mathematics. A numerical number representing each data point's level of participation in the clustering center is the clustering outcome. Assume that dataset I = {I1, I2,……,In} is the collection of k data, with f features in each sample Ix. M = {M1, M2,……,MC} is the matrix with the center, and the k sample is divided into the group of C fuzzy. The objective function is determined as,

The KDD dataset is divided by the weighted FCM algorithm, an unsupervised learning method based on objective function optimization that iteratively optimizes the objective function. Fuzzy mathematics is used in the FCM technique to study the clustering problem. The degree of participation of each data point in the clustering center is described by the clustering outcome, which is a numerical value.

The accumulation of k data, where each sample Ix has f features, is represented by dataset I = {I1, I2,……,In}. The matrix with center M = {M1, M2,……,MC} is used to separate the k sample into the group of C fuzzy. The objective function is established as follows:

$$OBF(L, M) = \sum_{Cx=1} \sum_{ky=1} (lxy) v(dxy)2 \quad (3)$$

where lxy is the membership between Mi and Iy, and L = lxy is a k × c dimension of the membership matrix. The Euclidean distance between sample y and cluster center point x is dxy = Iy − Mx. The

fuzzy exponent in the algorithm, v (v > 1), is usually given the value 2. It must fulfill the following requirements:

The enormous volume of data produced is separated into clusters and organized according to DPWFCM before being forwarded to the map reduce phase. Hadoop's MapReduce architecture is used to create apps that can handle enormous amounts of data on enormous clusters. It is also acknowledged as a programming technique for dividing up big datasets among several machines. Data can be stored in a distributed format with this software. It makes large-scale

The S-Box table for the Exclusive EnhancedWAKE technique is flexible and different for each round. The first step in the key creation process is to split 128-bit keys into 8 subkeys of 16 bits each, rotate the key to 25 bits to the left, and then split it into 8 subkeys again. The 64-bit plaintext is split up into four 16-bit sub-blocks in the first encryption round. In addition to using six bits of 16-bit sub-keys at each iteration, four sub-blocks are XOR-aligned, joined, and multiplied by one another. In order to achieve great generalization capabilities, the WSVM starts by creating a cost function and then aims to increase the range of divergence while minimizing the incidence of classification error. Unlike regular SVM, where the value of C is fixed and all training data points (noises and outliers) are treated uniformly during the training process, WSVM balances the penalty term to reduce the impact of minimally significant data points. The constrained-based optimization problem is expressed as

$$\sum C\ x=1\ lxy = 1y = 1, 2, 3, \ldots \ldots .n;$$
$$0 \leq lxy \leq 1\ x = 1, 2, 3, \ldots \ldots .C;$$
$$y = 1, 2, 3, \ldots \ldots .n;$$
$$0 < \sum k\ y=1\ lxy < kx = 1, 2, 3, \ldots \ldots . C;$$
$$Mx = (6)\ \sum k\ y=1\ (lxy)\ v\ Iy\ \sum k\ y=1\ (lxy)$$

The only obvious difference between SVM and the suggested WSVM is the upper constraints of Lagrange multipliers in the dual issue. The influence of the proposed WSVM for reducing outlier impact is only noticeable when the value of C is countably positive.

Similar to the well-known evolutionary strategy of genetic algorithms, scatter search is an evolutionary methodology that uses a combination method to integrate the attributes of two parent vectors to generate many children. It also uses expert systems to provide local solution tuning in an efficient manner. One crucial aspect of scatter search is the trade-off between combination strategies with exploration capabilities and augmentation mechanisms with exploitation capabilities.

The process of improvement begins once a solution is initially found through the development of a diversification technique. The generative technique is used to generate new elements and subsets, and the results are arranged according to the objective function. To update the progress, the combination and improvement procedures are used. The solution set's worst solutions are removed.

1 Result and discussion

The suggested method The Cumulative distribution function covered in this part has been utilized to create Exclusive Enhanced WAKE with Hadoop framework in MATLAB. The approach is applied to the KDD data set. The suggested efficient convolution approach for privacy-preserving across massive data using map reduction idea in a cloud computing environment was tested using the census-income KDD data set. The suggested approach's performance is assessed using the following metrics: data uploading time, accuracy, encryption time, decryption time, processing time, and categorization time. The suggested approach is compared to current approaches. The dataset description for the KDD data collection is displayed in Table 1.

1.1 Data uploading time

Data in the cloud system is handled with the assistance of hadoop mechanism and the data uploading is effective with the hadoop.

KDD set	Attack	Norma l	Total

Training Set	3,925,650	972,78 4,898,431 1
Testing Set	250,436	60,591 311,027

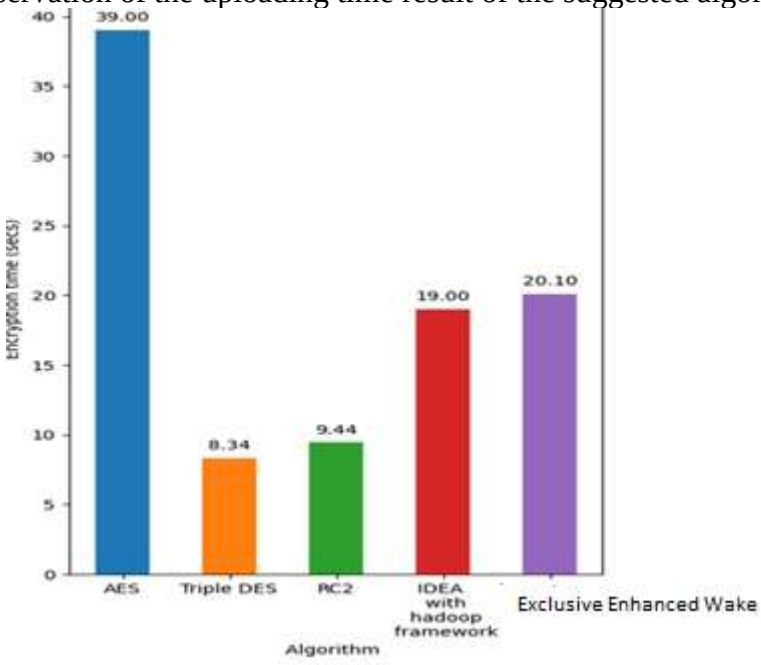
Table 1- KDD Data set

Algorithm	Data uploading time (seconds)
AES	40
Triple DES	35
RC2	38
IDEA with hadoop framework	11.8
Exclusive Exclusive EnhancedWAKE	10.3

Table 2 – Comparison of Data Uploading Time

Hadoop speeds up data uploading, and data security measures can impede the efficient technique's ability to handle the uploading. The time it takes for each algorithm to upload the specified amount of data is displayed in Table 2 and Figure 1.

Based on the observation of the uploading time result of the suggested algorithm, it is



Data Uploading Time

1.2 Encryption time

Table 3 and Fig. 2 show the time it took to encrypt the data. Based on the observation of the proposed algorithm's encryption time, it is determined that the time consumption of the proposed methodology is relatively low when compared with the existing approaches, namely AES, triple DES, RC2, and IDEA. The encryption time is used to calculate an encryption method's bandwidth

and specifies the rate at which data is encrypted.

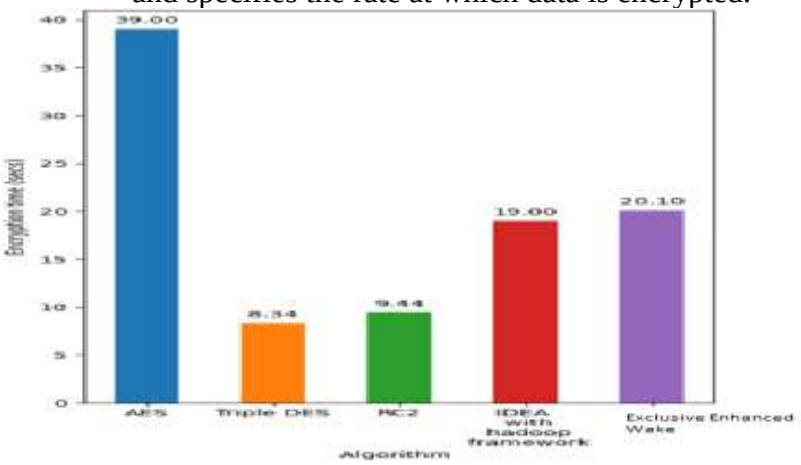


Fig 2-Data Encryption Time

Table 3 Comparison of encryption time

Algorithm	Encryption time (seconds)
AES	39
Triple DES	8.3377
RC2	9.435
IDEA with hadoop framework	19
Exclusive Exclusive EnhancedWAKE	20.1

1.3 Decryption time

The decryption key is used to change the modified files. Table 4 notes and provides the amount of time needed to decrypt the encrypted data. The minimum decryption time is used to gauge how effective the proposed algorithm is. The time required to decode the data is displayed in Table 4 and Fig. 3.

Table 4 Comparison of Decryption time

Algorithm	Decryption time (seconds)
AES	0.12
Triple DES	0.23
RC2	0.20
IDEA with hadoop framework	17.4
Exclusive Exclusive EnhancedWAKE	18.6

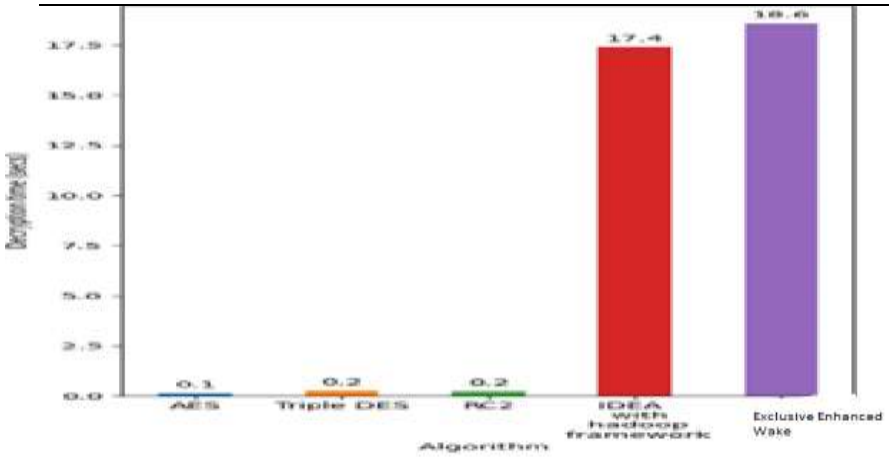


Fig 3.-Data Decryption Time

When compared to the current methods, which include AES, triple DES, RC2, and IDEA, the suggested methodology's time consumption is quite low, according to the results of the proposed algorithm's decryption time. Fig. 3 shows the time required to decode the data.

1.1 Data processing time

The decryption key is used to change the modified files. Table 4 notes and provides the amount of time needed to decrypt the encrypted data. The effectiveness of the proposed algorithmthe amount of memory pages required to complete the cryptographic method's encryption and decryption, as well as the execution time. Data processing time is the amount of time needed to process the necessary and private data using algorithms that carry out complete cryptography. The processing time is displayed in Fig. 4 and Table 5.

Table 5 Comparison of data Processing time

File size (MB)	100	300	500	1000	2000
GMPLS/MPLS networks	58	76	100	135	190
IDEA with hadoop framework	17.04	51.12	85.2	170.4	340.8
Exclusive Enhanced WAKE	18.6	53	86	272	434

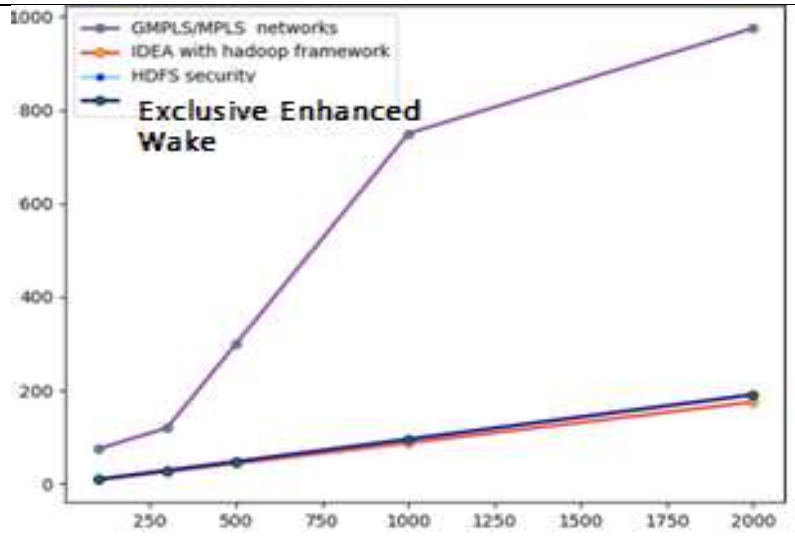


Fig. 4- data processing time

1.1 Data classification time

Data categorization time is the amount of time needed to categorize various attack types and standard action. Table 6 and Fig. 5 show the time required to classify the data. The results of the data classification time of the suggested algorithm show that, in comparison to the current methods, GMPL and IDEA, the time consumption of the suggested methodology is relatively low. Fig. 5 shows the amount of time needed to classify the data.

GMPLS/MPLS networks	75	120	300	750	975
IDEA with Hadoop framework	8.77	26.31	43.85	87.7	175.4
Enhanced WAKE	10	29	48	96	191

Table 6:Data Classification Time

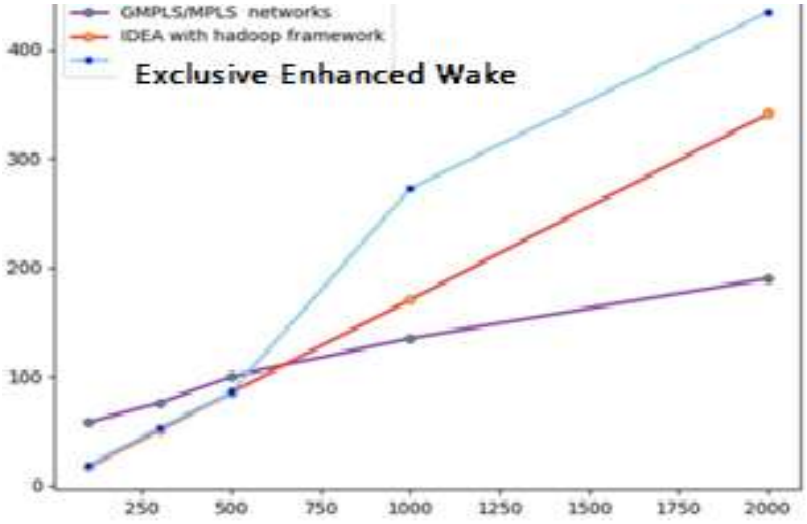


Fig 5:Data Classification Time

The accuracy of the classification indicates how efficient the suggested categorization method is. The effectiveness is contrasted with the current methods, which include artificial neural networks (ANN) [20], genetic algorithms (GA) [13], support vector machines (SVM), and naïve bayes (NB) [25]. The results are displayed in Fig. 6 and Table 7.
It is evident from the observation that the suggested

Table 7.Comparison of Accuracy, Precision and Recall

Algorithm	Accuracy %	Precision %	Recall %
SVM	88	86	86.4
NB	79	77	78.8
GA	83	85	89
ANN	90	91	91.5
AWSVM-CSS	96.1	97	96.9

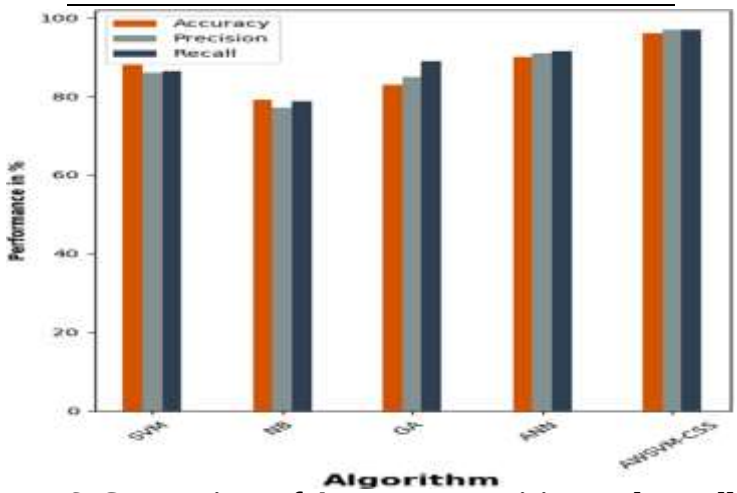


Fig. 6**. Comparison of Accuracy, Precision and Recall**

Conclusion

The big data security technique was developed and introduced based on the big data classification technique. Before any data is transferred across cloud nodes using this strategy, only the files identified as secret by the classification process should be protected. The efficiency gained by the specified approach strategies is displayed in the graphical and tabular forms. It is measured how long

it takes to classify large amounts of data. The categorization technique shows a noticeable improvement in speed by minimizing the need for repeated encryption and decryption processes for both public and mixed files. The big data-based Hadoop security method is evaluated in terms of reaction time and delay time. The efficiency of the suggested strategy is demonstrated by its 96.9% accuracy rate. The assessment

References

1. AbdElminaam DS, Abdual-Kader HM, Hadhoud MM (2010) Evaluating the performance of symmet- ric encryption algorithms. Int J Netw Secur 10(3):216–222
2. Almorsy M, Grundy J, Müller I (2016) An analysis of the cloud computing security problem. arXiv preprint arXiv:1609.01107
3. Chandra MA, Bedi SS (2021) Survey on SVM and their application in image classification. Int J Inf Technol 13(5):1–11
4. Del Rosal E, Kumar S (2017) A fast FPGA implementation for triple DES encryption scheme. Circuits Syst 8(09):237
5. Ding J, He X, Yuan J, Jiang B (2018) Automatic clustering based on density peak detection using gen- eralized extreme value distribution. Soft Comput 22(9):2777–2796
6. George Amalarethinam DI, Rajakumari S (2019) A survey on security challenges in cloud computing
7. Gontumukkala SST, Godavarthi YSV, Gonugunta BRRT, Subramani R, Murali K (2021) Analysis of image classification using SVM. In: 2021 12th International Conference on Computing Communica- tion and Networking Technologies (ICCCNT). IEEE, pp 01–06
8. Gordon A (2016) The hybrid cloud security professional. IEEE Cloud Comput 3(1):82–86
9. Hassan H, Nasir MHM, Khairudin N, Adon I (2017) Factors influencing cloud computing adoption in small medium enterprises. J Inf Commun Technol 16(1):21–41
10. Hentschel R, Leyh C, Petznick A (2018) Current cloud challenges in Germany: the perspective of cloud service providers. J Cloud Comput 7(1):1–12
11. Hou J, Pelillo M (2016) A new density kernel in density peak-based clustering. In: 2016 23rd Interna- tional Conference on Pattern Recognition (ICPR). IEEE, pp 468–473
12. Huang S, Guo Y, Yang N, Zha S, Liu D, Fang W (2021) A weighted fuzzy C-means clustering method with density peak for anomaly detection in IoT-enabled manufacturing process. J Intell Manuf 32(7):1845–1861
13. Jiang L, Zhang L, Li C, Wu J (2018) A correlation-based feature weighting filter for Naive Bayes. IEEE Trans Knowl Data Eng 31(2):201–213
14. Jimenez JM, Diaz JR, Lloret J, Romero O (2018) MHCP: multimedia hybrid cloud computing protocol and architecture for mobile devices. IEEE Netw 33(1):106–112
15. Kalra M, Tyagi S, Kumar V, Kaur M, Mashwani WK, Shah H, Shah K (2021) A comprehensive review on scatter search: techniques, applications, and challenges. Math Probl Eng 2021, pp 1–21
16. Kashani MH, Madanipour M, Nikravan M, Asghari P, Mahdipour E (2021) A systematic review of IoT in healthcare: applications, techniques, and trends. J Netw Comput Appl 192:103164
17. Khan MA (2016) A survey of security issues for cloud computing. J Netw Comput Appl 71:11–29
18. Li K, Wu Y, Nan Y, Li P, Li Y (2019) Hierarchical multi-class classification in multimodal spacecraft data using DNN and weighted support vector machine. Neurocomputing 259:55–65
19. Lu CC, Tseng SY (2002) Integrated design of AES (Advanced Encryption Standard) encrypter and decrypter. In: Proceedings IEEE international conference on application-specific systems, architec- tures, and processors. IEEE, pp 277–285
20. Mirjalili S (2019) Genetic algorithm. In: Evolutionary algorithms and neural networks. Springer, Cham, pp 43–55
21. Nurdiyanto H, Rahim R, Ahmar AS, Syahril M, Dahria M, Ahmad H (2018) Secure a transaction activ- ity with base64 algorithm and word auto key encryption algorithm. J Phys: Conf Ser 1028(1):012053). IOP Publishing
22. Ramachandra G, Iftikhar M, Khan FA (2017) A comprehensive survey on security in cloud computing. Procedia Comput Sci 110:465–472

23. Raut RD, Gardas BB, Jha MK, Priyadarshinee P (2017) Examining the critical success factors of cloud computing adoption in the MSMEs by using ISM model. J High Technol Manag Res 28(2):125–141
24. Razzaghi T, Roderick O, Safro I, Marko N (2016) Multilevel weighted support vector machine for classification on healthcare data with missing values. PLoS ONE 11(5):e0155119
25. Suthaharan S (2016) Support vector machine. In: Machine learning models and algorithms for big data classification. Springer, Boston, pp 207–235
26. Tariq MI, Memon NA, Ahmed S, Tayyaba S, Mushtaq MT, Mian NA, ... Ashraf MW (2020) A review of deep learning security and privacy defensive techniques. Mob Inf Syst 2020:6535834
27. Tharini VJ, Shivakumar BL (2022) High-utility itemset mining: fundamentals, properties, techniques and research scope. In: Computational intelligence and data sciences. CRC Press, Boca Raton, pp 195–210
28. Tharini VJ, Vijayarani S (2020) IoT in healthcare: ecosystem, pillars, design challenges, applications, vulnerabilities, privacy, and security concerns. In: Incorporating the Internet of Things in healthcare applications and wearable devices. IGI Global, pp 1–22
29. Vijayarani S, Tharini VJ, Sivamathi C (2019) Business intelligence for evaluating the intangible benefits of on-shelf high utility itemset from the temporal transaction database. International Journal of Engineering and Advanced Technology (IJEAT) 8(6)
30. Walczak S (2018) Artificial neural networks. In: Encyclopedia of information science and technology, Fourth Edition. IGI Global, pp 120–131