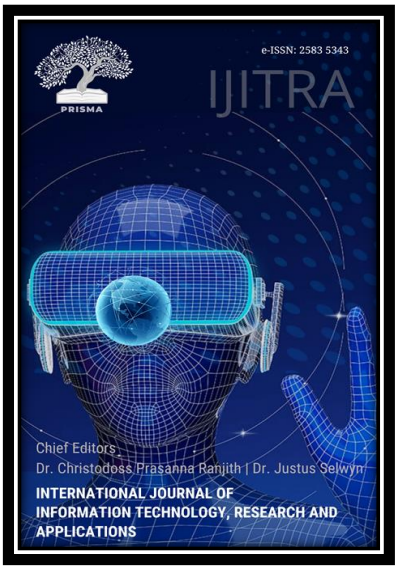


# International Journal of Information Technology, Research and Applications (IJITRA)

**Mahesh T R, Vivek V, Vinoth Kumar V.**
**Implementation of Machine Learning-Based Data Mining Techniques for IDS. International Journal of Information Technology, Research and Applications, 2(1), 07-13.**

**ISSN: 2583-5343**

**DOI: 10.59461/ijitra.v2i1.23**

The online version of this article can be found at:
https://www.ijitra.com/index.php/ijitra/issue/archive

Published by:
PRISMA Publications



**International Journal of Information Technology, Research and Applications (IJITRA)** is a journal that publishes articles which contribute new theoretical results in all the areas of Computer Science, Communication Network and Information Technology. Research paper and articles on Big Data, Machine Learning, IOT, Blockchain, Network Security, Optical Integrated Circuits, and Artificial Intelligence are in prime position.

PRISMA PUBLICATIONS

https://www.prismapublications.com/

# Implementation of Machine Learning-Based Data Mining Techniques for IDS

***Mahesh T R[1], V Vivek[1], Vinoth Kumar[1]***
[1] Department of Computer Science and Engineering, Faculty of Engineering and Technology,
JAIN (Deemed-to-be University), Bangalore, India

**Article Info**

***Article history:***

Received September 26, 2022
Revised October 29, 2022
Accepted December 20, 2022

***Keywords:***

Data Mining
Intrusion Detection System
Machine Learning

**ABSTRACT**

The internet is essential for ongoing contact in the modern world, yet its effectiveness might lessen the effect known as intrusions. Any action that negatively affects the targeted system is considered an intrusion. Network security has grown to be a major issue as a result of the Internet's rapid expansion. The Network Intrusion Detection System (IDS), which is widely used, is the primary security defensive mechanism against such hostile assaults. Data mining and machine learning technologies have been extensively employed in network intrusion detection and prevention systems to extract user behaviour patterns from network traffic data. Association rules and sequence rules are the main foundations of data mining used for intrusion detection. Given the Auto encoder algorithm's traditional method's bottleneck of frequent itemsets mining, we provide a Length-Decreasing Support to Identify Intrusion based on Data Mining, which is an upgraded Data Mining Techniques based on Machine Learning for IDS. Based on test results, it appears that the suggested strategy is successful.



***Corresponding Author:***

Mahesh T R
Department of Computer Science and Engineering,
JAIN(Deemed-to-be University),
Bengaluru, India.
Email: trmahesh.1978@mail.com

## I. INTRODUCTION

A piece of software called an intrusion detection system (IDS) keeps an eye on and protects a network from intruders. The rapid development of Internet-based technologies has resulted in the development of numerous application aspects for CNs. A few of the LAN & WAN applications that have become more commonplace recently are those related to business, finance, industry, security, and healthcare [1-4]. Area networks are an enticing target for theft because of all these uses, endangering the neighborhood. The internal systems of a business are used by malicious users or hackers to gather data, take advantage of software flaws, and exploit operational issues before bringing the system back to default. As the Internet grows more common in society, new things like viruses and worms are introduced [5-7]. Due of its lethal nature, users can use it to delete unencrypted text and passwords.

Users need security as a result to safeguard their systems from intrusions. A well-known security technique for protecting both private and public networks is firewall technology. IDS is used by insurance companies, medical apps, credit card fraud, and system-related operations [8-10]. The goal of an IDS is to detect malicious traffic. This is accomplished by the IDS by monitoring all incoming and outgoing traffic. A few methods can be used to implement an IDS. The most well-liked two of these are: spotting irregularities The foundation of this strategy is the detection of traffic irregularities. Calculated is the observed traffic's deviation from the typical profile. There have been numerous implementations of this strategy that are based on measures for calculating traffic profile deviance.

Numerous researchers used various techniques, such as early-stage screening, and created novel ways for the early prediction of cancer therapy outcomes [11-14]. In the realm of medicine, cutting-edge technologies are used, and the medical research community has access to vast volumes of cancer data that have been gathered. Machine learning techniques are now a well-liked tool for medical researchers. Numerous machine learning techniques, including feature selection and classification, are frequently used in cancer detection. To find patterns and relationships in large datasets, machine learning techniques are applied [15-18].

Signature/Misuse Detection: This method looks for patterns and indicators of previously known attacks in network traffic. The signatures of well-known attacks are often kept in a regularly updated database. This method of handling intrusion detection is comparable to how anti-virus software functions. An alarm for theft has been activated, according to IDS [19-21]. A home can be safeguarded against theft, for instance, using a lock system. If someone tries to break into the house by breaking the lock system, the thief alarm detects the lock being broken and sounds an alarm. Additionally, firewalls do a superb job of screening incoming Internet traffic. A typical Intrusion detection system is shown in figure (1).

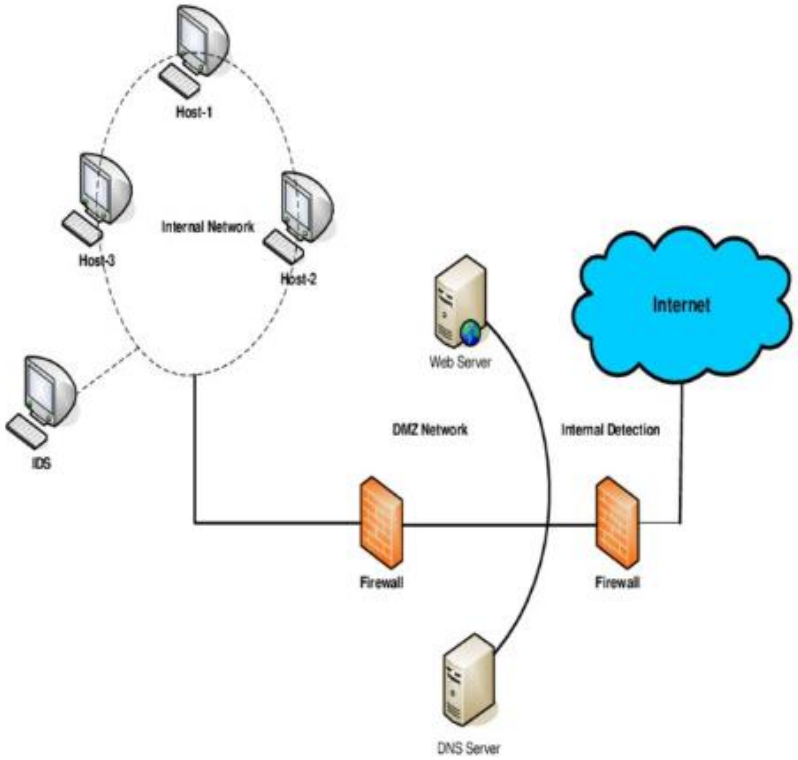


**Figure 1.** Intrusion detection system

For instance, external operators can access the intranet via a modem linked to the company's private network; however, a firewall does not protect this form of access. To identify and stop vulnerable traffic, an intrusion prevention system (IPS) analyses network traffic. Network (NIPS) and host defense systems are the two categories of defense systems (HIPS). These systems monitor network activity and take automated security measures to protect the system and network. There are numerous false positives and suggestions in the IPS issue. Events that trigger an IDS alarm but don't lead to a successful attack are known as false positives. An incident that doesn't trigger an alarm during an assault is known as a false negative. Single points of failure, altered signatures, and encoded traffic are a few instances of inline functionality that could be problematic. Using IDS, the system's performance is evaluated [22-25].

## II. RELATED WORKS

To progress the field of machine learning and employ it in several study fields, including healthcare, it is necessary to create newer algorithms. According to a recent study, machine learning can be used to treat cancer [26-27]. A new era of individualized medicine with swift and sophisticated data analysis, previously unreachable, is beginning with the use of machine learning and AI techniques in basic and translational cancer research. Countless data sets and machine learning algorithms aid in the diagnosis, treatment, and prognosis of cancer, among other aspects of the fight against disease. Machine learning makes it possible to tailor the therapy to the patient, which would not be possible without it.

In this study [28], a method for using an orthopantomogram to detect oral tumors is proposed. To maintain these edge characteristics as well as the conspicuous watershed on images, which causes over segmentation despite being pre-processed, a novel mathematical morphological watershed approach is suggested. Marker controlled watershed segmentation is used to segment tumors to prevent over segmentation. In this paper [29] a hybrid model is put forth that consists of two stages, the first of which uses the ReliefF-GA feature selection method to identify the best feature of the subset and the second of which uses the ANFIS classification to categories patient survival after a specific number of years since diagnosis.

Two datasets of oral cancer with clinicopathologic and genomic markers each were used to test the suggested predictive model. It has been tested that the suggested model performs better when both types of datasets are used, along with additional techniques like logistic regression, support vector machines, and artificial neural networks [30-34].

The detection rate of intrusive attacks can be improved with this technology when compared to earlier single learning model strategies. Thanks to a parallel training method and huge data methodologies, the model building time of BDHDLS is drastically reduced when multiple machines are deployed. In this [35-38], they deployed a machine learning-based intrusion detection system (Random Forest) for CSE-CIC-IDS-2018, and it delivered a remarkable score of 99 percent accuracy. The NSL-KDD data set was used by the authors of this paper to conduct a thorough review of several studies pertinent to machine learning-based IDS. They suggested a generic process flow for anomaly-based IDS and spoke about its elements in relation to earlier studies. then offered a few intriguing ideas for future research [39-41]. This article explores the potential integration of machine learning and data mining techniques with intrusion detection technology for cyber security. Phishing detection will eventually be automated and transformed into an artificial intelligence task since neural networks, which are the basis of intelligence, are constructed on multi-layer perceptrons.

## III. PROPOSED METHODOLOGY

The study of self-improving computer algorithms is referred to as machine learning. Applications range from information filtering systems that automatically learn consumers' preferences to data mining techniques that discover general principles in large data sets [42]. Contrary to statistical methods, machine learning approaches are perfect for learning patterns without any prior knowledge of what such patterns might be. Clustering and classification problems are the two most prevalent ones in machine learning. Both problems have been addressed through strategies that have been applied to IDSs. 1) Classification methods: The objective of a classification task in machine learning is to categorize every instance of a dataset. An IDS that employs classification tries to classify all traffic as either benign or malicious. The objective is to lessen the number of false positives (traffic deemed damaging but not malicious) and false negatives (classification of malicious traffic as normal).

In the past, a hybrid clustering-based approach was utilized, where the right number of clusters was decided upon first, followed by clustering. The data was grouped using the Auto encoder technique, and the genetic method was used to estimate the ideal number of clusters [43]. Due to the limitations of the prior study, we chose to use more advanced clustering approaches that produce superior outcomes to the earlier work. We use the information-gathering strategy to choose attributes in the provided method. The data was then grouped using the Fuzzy C-means approach, with the optimal number of clusters being determined using DE.

DE is a novel heuristic approach that has three benefits: it is rapid, it allows for control parameters, and it discovers the original global minimum regardless of the initial parameter values. The DE algorithm uses crossover, mutation, and selection operators like a genetic algorithm. It is a population-based algorithm. To function, the Autoencoder algorithm adheres to a set of principles. Each data point has a cluster center associated with it based on the distance between the cluster core and the data point. You may find out more around about the cluster center and its participants. Each data point's membership must unambiguously equal one. After using the formulation as indicated in equations (1) and (2), update each periodic membership and cluster center.

$u_{ij}=1/\sum_{k=1}^{c}(d_{ij}/d_{ik})^{(\frac{2}{m}-1)}$ (1)

$v_j=(\sum_{i=1}^{n}(u_{ij})^m x_i)/ (\sum_{i=1}^{n}(u_{ij})^m)$ , $\forall_j=1,2 \ldots.c$ (2)

Where n→ *no of data points*
$v_j \longrightarrow jth\ cluster\ center$
M → fussiness index
c→ no of cluster centres
$u_{ij}$ →membership of ith data to jth cluster center
$d_{ij}$ −→Euclidean distance amongst ith data & jth cluster center.

Key objective of fuzzy c-means algorithm is to minimize using the equation (3).

$J(U,V) = \sum_{i=1}^{n}\sum_{i=1}^{c}(u_{ij})^m \quad ||x_i - v_j||^2$ (3)

Where $||x_i - v_j||^2$Euclidean distance amongst ith data & jth cluster center.
The dataflow diagram of IDS is displayed in Fig.2. The flow diagram shows he steps involved in the

implementation of this research work. IDS can be secluded in following factors: Dataset, Feature Selection, Training Phase, Testing Phase and Classifier.

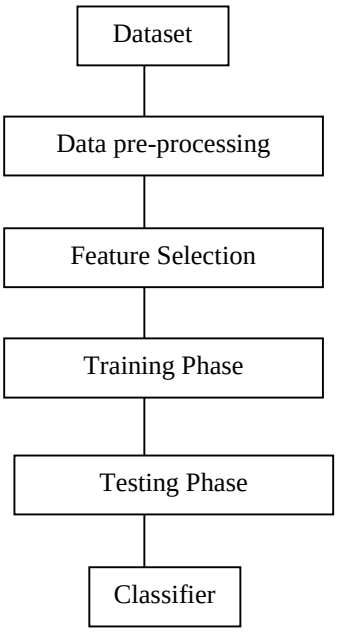


Figure 2. Process Diagram

Utilizing classifiers, intrusion detection systems are assessed. Whether the result of the algorithm is accurate or not. Our classifier utilizes ID mapping to guarantee the output of the class is accurate. After removing the attribute, the dataset is compressed into seven attribute datasets, one of which is the attribute ID number. K refers to the ID number in the CSE-CIC-IDS2018 dataset for the same sample. The ID number of the reference output is used to cross-reference and verify accuracy. A positive no is referred to as a "true positive" (TP). tuples from the classifier that have been correctly labelled. False positive (FP): This term describes the quantity of negative tuples that classifiers inadvertently identify as positive. False Negatives are positive tuples that were mistakenly classified as negatives (FN).

The proportion of true positives to false positives is how precision is measured.

$$Precision = \frac{TP}{FP} \quad (4)$$

The ratio of true positives to the total of false positives and false negatives is known as recall.

$$Recall = \frac{TP}{(FP + FN)} \quad (5)$$

The overall accuracy of the classifier is its accuracy.

$\frac{Precision}{REcall}$ = Accuracy (6)

Table 1. Comparison of the accuracy of the current approach, the proposed approach, and the AWS dataset CSE-CIC-IDS2018

| | **Precision** | **Recall** | **Accuracy** |
|---|---|---|---|
| Existing | 79.15 | 81.27 | 80.24 |
| Proposed | 92.29 | 93.17 | 93.12 |

Table 1 compares proposed and ongoing research projects including the Auto encoder method. The comparison demonstrates that the proposed work has better precision and recall as well as better accuracy than the prior effort.

## IV. CONCLUSION

Crimes involving intrusions are on the rise. Therefore, compared to intrusion detection systems that make use of conventional clustering techniques, the best intrusion detection system must be found. In this study, we developed an intrusion detection system that does not have a predetermined group number (k) and instead uses the algorithm to identify the type of intrusion. The fitness function is used to calculate the ideal value of

K, which facilitates the efficient creation of optimized clusters and improves the effectiveness of type of attack detection. as opposed to intrusion defenses.

## V. REFERENCES


[1] H. K. Shashikala, T. R. Mahesh, V. Vivek, M. G. Sindhu, C. Saravanan and T. Z. Baig, "Early Detection of Spondylosis using Point-Based Image Processing Techniques," 2021 International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT), 2021, pp. 655-659, doi: 10.1109/RTEICT52294.2021.9573604.

[2] Gowramma, G. S., Mahesh, T. R., & Gowda, G. (2017). An automatic system for IVF data classification by utilizing multilayer perceptron algorithm. *ICCTEST-2017*, *2*, 667-672.

[3] M. R. Sarveshvar, A. Gogoi, A. K. Chaubey, S. Rohit and T. R. Mahesh, "Performance of different Machine Learning Techniques for the Prediction of Heart Diseases," 2021 International Conference on Forensics, Analytics, Big Data, Security (FABS), 2021, pp. 1-4, doi: 10.1109/FABS52071.2021.9702566.

[4] Dharahas, R. T., & Mahesh, T. R. (2021). A Pragmatic Approach for Detecting Brain Tumors Using Machine Learning Algorithms. *BIOSCIENCE BIOTECHNOLOGY RESEARCH COMMUNICATIONS Special Issue*, *14*(11).

[5] Pinaki, G., & Mahesh, T. R. (2015). Smart city: Concept and challenges. *International Journal on Advances in Engineering Technology and Science*, *1*(1).

[6] P. Chaitanya Reddy, R. M. S. Chandra, P. Vadiraj, M. Ayyappa Reddy, T. R. Mahesh and G. Sindhu Madhuri, "Detection of Plant Leaf-based Diseases Using Machine Learning Approach," 2021 IEEE International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS), 2021, pp. 1-4, doi: 10.1109/CSITSS54238.2021.9683020.

[7] R. Pasumarty, R. Praveen and M. T. R, "The Future of AI-enabled servers in the cloud- A Survey," 2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2021, pp. 578-583, doi: 10.1109/I-SMAC52330.2021.9640925.

[8] Mahesh, T. R., & Naik, D. M. K. (2020). Analysis of Academic Performance in massive Open Online Courses (Moocs) Using Process mining. *International Journal of Computer Trends and Technology*, *68*(12), 21-25.

[9] Mahesh, T.R., Vinoth Kumar, V., Vivek, V. *et al.* Early predictive model for breast cancer classification using blended ensemble learning. *Int J Syst Assur Eng Manag* (2022). https://doi.org/10.1007/s13198-022-01696-0

[10] T. R. Mahesh, V. Vivek, V. V. Kumar, R. Natarajan, S. Sathya and S. Kanimozhi, "A Comparative Performance Analysis of Machine Learning Approaches for the Early Prediction of Diabetes Disease," 2022 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), 2022, pp. 1-6, doi: 10.1109/ACCAI53970.2022.9752543.

[11] P. Shrestha, A. Singh, R. Garg, I. Sarraf, T. R. Mahesh and G. Sindhu Madhuri, "Early Stage Detection of Scoliosis Using Machine Learning Algorithms," 2021 International Conference on Forensics, Analytics, Big Data, Security (FABS), 2021, pp. 1-4, doi: 10.1109/FABS52071.2021.9702699.

[12] K. K. Jha, R. Jha, A. K. Jha, M. A. M. Hassan, S. K. Yadav and T. Mahesh, "A Brief Comparison On Machine Learning Algorithms Based On Various Applications: A Comprehensive Survey," 2021 IEEE International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS), 2021, pp. 1-5, doi: 10.1109/CSITSS54238.2021.9683524.

[13] S C, Asha and S C, Asha and TR, Mahesh and V, Vivek and M B, Suresh, The Importance of Teacher's Mental Health and Wellness for Quality Learning in Classrooms during COVID-19 Pandemic (April 30, 2021). Available at SSRN: https://ssrn.com/abstract=3837304 or http://dx.doi.org/10.2139/ssrn.3837304

[14] Mahesh, T. R., & Naik, D. M. K. (2021). A Comprehensive Review of Behavioral Customer Segmentation For A Better Understanding. *International Journal of Computer Science and Engineering*, *8*(1), 1-4.

[15] Tarun, R. R., Sahana, J. S., Sadvik, B. S., Shashank, S., & Mahesh, T. R. (2019). Context based Sentiment Analysis of Twitter using Hadoop Framework. *International Journal of Computer Science and Mobile Computing*, *8*(5), 193-202.

[16] Subashini, S., & Mahesh, T. R. Web Mining: Prominent Applications and Future Directions. *International Journal of Computer Science and Information Technology & Security*, 825-830.

[17] S. Roopashree, J. Anitha, T.R. Mahesh, V. Vinoth Kumar, Wattana Viriyasitavat, Amandeep Kaur, An IoT based authentication system for therapeutic herbs measured by local descriptors using machine learning approach, Measurement, Volume 200, 2022, 111484, ISSN 0263-2241, https://doi.org/10.1016/j.measurement.2022.111484.

[18] Mahesh, T. R., Kumar, D., Vinoth Kumar, V., Asghar, J., Mekcha Bazezew, B., Natarajan, R., & Vivek, V. (2022). Blended Ensemble Learning Prediction Model for Strengthening Diagnosis and Treatment of Chronic Diabetes Disease. *Computational Intelligence and Neuroscience*, *2022*.

[19] A. Srivastava, V. V. Kumar, M. T. R and V. Vivek, "Automated Prediction of Liver Disease using Machine Learning (ML) Algorithms," 2022 Second International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), 2022, pp. 1-4, doi: 10.1109/ICAECT54875.2022.9808059.

[20] S. Surana, K. Pathak, M. Gagnani, V. Shrivastava, M. T. R and S. Madhuri G, "Text Extraction and Detection from Images using Machine Learning Techniques: A Research Review," 2022 International Conference on Electronics and Renewable Systems (ICEARS), 2022, pp. 1201-1207, doi: 10.1109/ICEARS53579.2022.9752274.

[21] Mahesh, T. R., Krishna, G. V., Sathwik, P., Chowdary, V. A., & Hemchand, G. (2022). Providing Voice to Susceptible Children: Depression and Anxiety Detected with the Help of Machine Learning. In *Integrated Emerging Methods of Artificial Intelligence & Cloud Computing* (pp. 444-450). Springer, Cham.

[22] Mahesh, T. R., Ram, M. S., Ram, N., Gowtham, A., & Swamy, T. V. (2022). Real-Time Eye Blinking for Password Authentication. In *Integrated Emerging Methods of Artificial Intelligence & Cloud Computing* (pp. 428-434). Springer, Cham.

[23] K. K. Jha, A. K. Jha, K. Rathore and T. R. Mahesh, "Forecasting of Heart Diseases in Early Stages Using Machine Learning Approaches," 2021 International Conference on Forensics, Analytics, Big Data, Security (FABS), 2021, pp. 1-5, doi: 10.1109/FABS52071.2021.9702665.

[24] TR, Mahesh and V, Vivek, Recommendation Systems: The Different Filtering Techniques, Challenges and Review Ways to Measure the Recommender System (April 14, 2021). Available at SSRN: https://ssrn.com/abstract=3826124 or http://dx.doi.org/10.2139/ssrn.3826124

[25] V. Vivek; T. R. Mahesh; C. Saravanan; K. Vinay Kumar, "5 A Novel Technique for User Decision Prediction and Assistance Using Machine Learning and NLP: A Model to Transform the E-commerce System," in Big Data Management in Sensing: Applications in AI and IoT , River Publishers, 2021, pp.61-76.

[26] G. Sindhu Madhuri; T. R. Mahesh; V. Vivek, "7 A Novel Approach for Automatic Brain Tumor Detection Using Machine Learning Algorithms," in Big Data Management in Sensing: Applications in AI and IoT , River Publishers, 2021, pp.87-102.

[27] G. Sindhu Madhuri; T. R. Mahesh; V. Vivek, "7 A Novel Approach for Automatic Brain Tumor Detection Using Machine Learning Algorithms," in Big Data Management in Sensing: Applications in AI and IoT , River Publishers, 2021, pp.87-102.

[28] V, Vivek and TR, Mahesh and Das, Shilpa, A Security Framework for Application Instance Data to Enhance the Privacy in Decentralized Cloud Environment (October 21, 2020). Available at SSRN: https://ssrn.com/abstract=3715953

[29] V, Vivek and TR, Mahesh and Das, Shilpa, A Security Framework for Application Instance Data to Enhance the Privacy in Decentralized Cloud Environment (October 21, 2020). Available at SSRN: https://ssrn.com/abstract=3715953

[30] Ghosh, P., & Mahesh, T. R. (2019). Untraceable privacy-preserving authentication protocol for RFID tag using salted hash algorithm. International Journal of Advanced Intelligence Paradigms, 13(1-2), 193-209.

[31] Aja-Fernández S., Curiale A. H., Vegas-Sánchez-Ferrero G., A local fuzzy thresholding methodology for multiregion image segmentation. Knowledge-Based Systems, 83:1–2. 2015.

[32] Patel K., Jha J., Brain tumor image segmentation using adaptive clustering and level set method, image, 9, 2014.
[33] Jaganathan P., Kuppuchamy R. A threshold fuzzy entropy based feature selection for medical database classification. Computers in Biology and Medicine, 43, 2222–2229, 2013.
[34] Sujan M., Alam N., Noman S. A., Islam M.J.. A segmentation based Automated System for Brain Tumor Detection. International Journal of Computer Applications, 153:41–9, 2016.
[35] Ilhan U., Ilhan A.. Brain tumor segmentation based on a new threshold approach, Procedia Computer Science, 120:580–587, 2017.
[36] DongjuLiu, JianYu, "Otsu method and K-means", 978-0-7695-3745-0/09, IEEE. DOI 10.1109/HIS.2009.74, 2009.
[37] Aimi Salihah Abdul-Nasir1 , Mohd Yusoff Mashor2 , Zeehaida Mohamed3 , " Colour Image Segmentation Approach for Detection of Malaria Parasites Using Various Colour Models and k-Means Clustering", Issue 1, Volume 10, January 2013.
[38] Juanying Xie, Shuai Jiang, "A simple and fast algorithm for global K-means clustering", 978-0-7695-3987-4/10, IEEE, 2010.
[39] Rafael C. Gonzalez & Richard E. Woods, "Digital Image Processing using Matlab", Third edition: Pearson education, 2005.
[40] Du, Cheng-Jin, and Da-Wen Sun. 2004. "Recent developments in the applications of image processing techniques for food quality evaluation." Review of. Trends in food science & technology, 15, (5):230-49.
[41] Du, Cheng-Jin, and Da-Wen Sun. 2004. "Recent developments in the applications of image processing techniques for food quality evaluation." Review of. Trends in food science & technology 15 (5):230-49.
[42] Shashikala H K, Sindhu Madhuri G, "Image pre-processing techniques for X-ray medical images: A Survey", International Journal of Creative Research Thoughts (IJCRT), ISSN: 2320-2882, Vol 9, Issue 1, January 2021.
[43] Sindhu Madhuri G, Shashikala H K, "Image Processing techniques for Detecting Extra Growth of Teeth in Medical Images", Solid State Technology, Vol 64, Issue 2, Jan 2021.