International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Comparative Study of Machine Learning Algorithms for Intrusion Detection

[1]Jyoti Dhanke, [2]R. N. Patil, [3]Indra Kumari, [4]Shiv Gupta, [5]Swati Hans, [6]Kaushal Kumar

Submitted: 11/09/2023 **Revised**: 21/10/2023 **Accepted**: 06/11/2023

Abstract: Researching Network Traffic Classification through Machine Learning is crucial given the expanding reach of the internet, enabling global information exchange. The implications of security breaches extend beyond individuals to impact entire organizations. Hence, discerning between malicious and non-malicious data on the network holds utmost significance. In this research, we perform an in-depth examination and contrast of seven distinct machine learning algorithms: Naïve Bayes, Logistic Regression, Support Vector Machine (SVM), Random Forest, C4.5, XGBoost, and k-Nearest Neighbors (KNN). These analyses are executed using Python's package module for seamless programmatic execution. The assessment encompasses metrics such as accuracy, precision, and recall, offering valuable insights into the performance of each algorithm.

Keywords: *Network Traffic Classification, Machine Learning, KNN, SVM*

1. Introduction

Presently, conventional intrusion detection models are inadequate in capturing the intricacies of the recent surge in cyber threats and incidents. The conventional approach involves manual network analysis or predefined fixed abnormal patterns, which may not effectively identify attacks within the system. With the growing network traffic facilitated by the internet, the ease of access to policy information complicates intrusion detection for network analysts. Automating the intrusion detection process demands dynamic and efficient methods capable of learning and detecting emerging intrusion types. This paper introduces highly adaptive and dynamic intrusion detection techniques tailored to handle substantial network traffic. The process of identifying intrusions involves three essential steps: defining and extracting features, defining and extracting rules, and applying these rules to detect intrusions within the dataset. These methods are designed to accommodate the specific requirements of diverse systems and networks. From the past three decade, many researchers propose different technique to classify the network traffic efficiently. we discuss some of the network classification technique that were used in the past by the researchers. Various techniques used to classify network traffic in the past are as Classification Based on Port Number and Classification Based on Payload.

2. Related Work

In [12] authors employed the KDD Cup 99 dataset to differentiate between normal and abnormal data. In another investigation conducted by Jamal H. Assi and Ahmed T. Sadiq (2017), the NSL-KDD dataset was utilized for the categorization of network attacks. This study involved the implementation of various classification methods, including Support Vector Machine (SVM), Decision Table (DT), C4.5, Bayesian Network, and Back Propagation Neural Network. Furthermore, diverse feature selection strategies were applied, such as Decision Table, Correlation-based feature selection (CFS), and Information Gain (IG). Notably, the C4.5 classification method with information gain feature selection demonstrated superior performance when compared to other algorithms [15][16][17].

Muhammad Shafiq, Xiangzhan Yu, Asif Ali Laghari, Lu Yao, and Nabin Kumar Karn formulated a proposal for a network traffic categorization framework. This proposal integrated four machine learning algorithms: C4.5, Support Vector Machine, BayesNet, and NaïveBayes, along with supervised learning techniques.

Dhanabal and Shantharajah (2015) used the NSL-KDD dataset to test a number of classification algorithms, concentrating on anomalies in network packets, such as SVM, Naïve Bayes, and J48.

Dataset Used:

In our research, we will employ the NSL-KDD dataset, an improved iteration of the original KDD Cup dataset. The KDD Cup dataset originated from the 1999

[1]*Bharati Vidyapeeth's College of Engineering, Lavale, Pune 412115, Maharashtra, India, jyoti.dhanke@bharatividyapeeth.edu*
[2]*Principal, Department of Mechanical Engineering, Bharati Vidyapeeth's College of Engineering Lavale Pune, India*
[3]*Korea Institute of Science and Technology Information, KISTI*
[3]*University of Science and Technology UST, South Korea*
[4]*IEC college of engineering and technology Greater Noida.*
[5]*Manav Rachna International Institute of Research & Studies, India Department of Computer Science & Engineering*
6Manav Rachna International Institite of Research and Studies Faridabad, Indiakaushalkumar.set@mriu.edu.in

International Knowledge Discovery and Data Mining tool competition, where the objective was to amass instances of network traffic. The primary aim of this competition was to create a predictive model capable of distinguishing between malicious and non-malicious data packets. The NSL-KDD dataset consists of 43 attributes for each instance, with 41 attributes providing information about the input traffic data, and the remaining 2 attributes indicating whether the data represents an attack or normal traffic. You can find detailed descriptions of these attributes in this paper [14].

The NSL-KDD dataset is split into separate files for both training and testing purposes. The training set comprises a total of 125,973 instances, which are classified into five distinct categories: Normal, DoS, Probe, R2L, and U2R [15].Likewise, the test set comprises 22,544 instances, also classified into these same five categories. Table 1 offers a detailed breakdown of the instance counts in each class.

Figures 1 and 2 depict the distribution of instances in the training and testing sets, correspondingly.

[20][21].

Table 1: No. of Instances in Each Class

Class	**Training for Set**	**% of Occurrence**	**Testing data Set**	**% of upcoming Occurrence**
Normal	67342	53.49%	9710	43.08%
DoS	45926	36.49%	7459	33.08%
Probe	11654	9.27%	2419	10.74%
R2L	993	0.78%	2880	12.22%
U2R	51	0.042%	65	0.89%

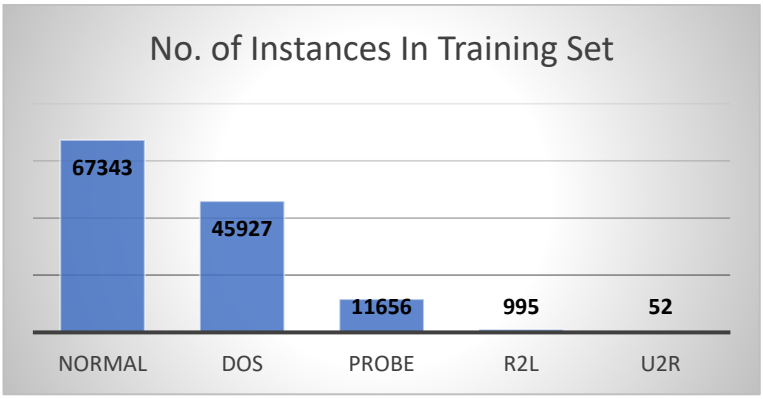

Fig 1: No. of Instances in Training Set

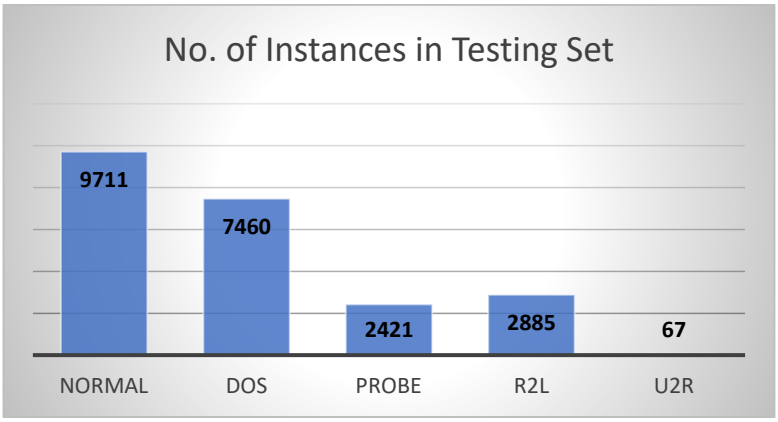

Fig 2: No. of Instances in Testing Set

Within the NSL-KDD using dataset, there are 4 distinct classes for attack:

i. DoS (Denial of Service): This attack involves overwhelming a network by inundating it with an excessive number of requests, rendering it inaccessible to its intended users. An example of this is SYN Flooding [16-20].

ii. Probe or Surveillance: In this attack, the attacker compromises the information of a remote computer, which can then be exploited for their malicious intentions. Port Scanning is one example of this type of attack.

iii. U2R (User to Root): In a U2R attack, a privilege to attempts to gain root privileges, thereby making the system vulnerable. An example of this is a Buffer Overflow attack [21-23].

iv. R2L (Remote to Local): In an R2L attack, the attacker endeavors for the gain access to the victim's system from a remote location, exploiting vulnerabilities in the system. Password Guessing is an example of this type of attack [22][23].

3. Proposed Methodology

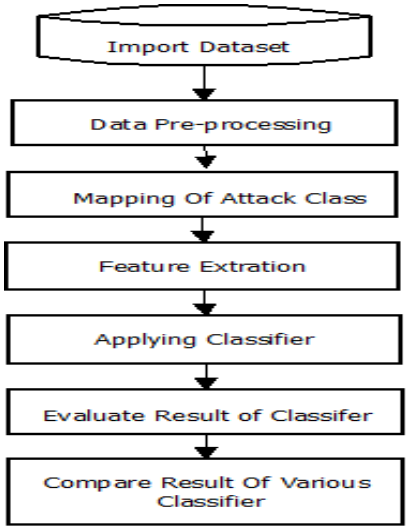

Fig 3: Methodology Used

i. The methodology adopted for the research consists of the following steps:

ii. **Data Pre-Processing:** A dataset is selected from the NSL_KDD dataset repository, followed by the application of pre-processing techniques to convert non-numeric attributes into numeric ones.

iii. **Mapping:** Various attacks are mapped to their corresponding attack classes.

iv. **Feature Selection:** This step involves applying a method for dimensionality reduction, with the utilization of Random Sampling to eliminate biased training.

v. **Applying Classifier:** Different machine learning classifiers are employed to classify the data.

vi. **Evaluating Performance Metrics**: The performance of the classifiers is assesing using the various parameters for classification accuracy, precision, and recall. The flowchart depicting the methodology utilized is presented in Figure 3.

Performance Evaluation and Experimental Analysis:

We used the analysis of the performance metrics as depicted in Table 2 below:

Table 2: Metrics for Performances

Metrics		**Actual Class**	
		A	**Not A**
Predicted Class	A	TRP	FLP
	Not A	FLN	TRN

The used performance metrics are as follows:

$$Accuracy = \frac{TRP + TRN}{TRP + FLP + FLN + TRN}$$

$$Recall = \frac{TRP}{TRP + FRN}$$

$$Precision = \frac{TRP}{TRP + FLP}$$

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Experimental Result: In our study, we employed the NSL-KDD dataset, selecting 20,000 instances from both the training and test sets. The frequency percentage of normal or attack class data is illustrated in Figure 4. Additionally, the presentation metrics of the seven classification utilized in the experiment, including accuracy, precision, and recall, are summarized in Table 3.

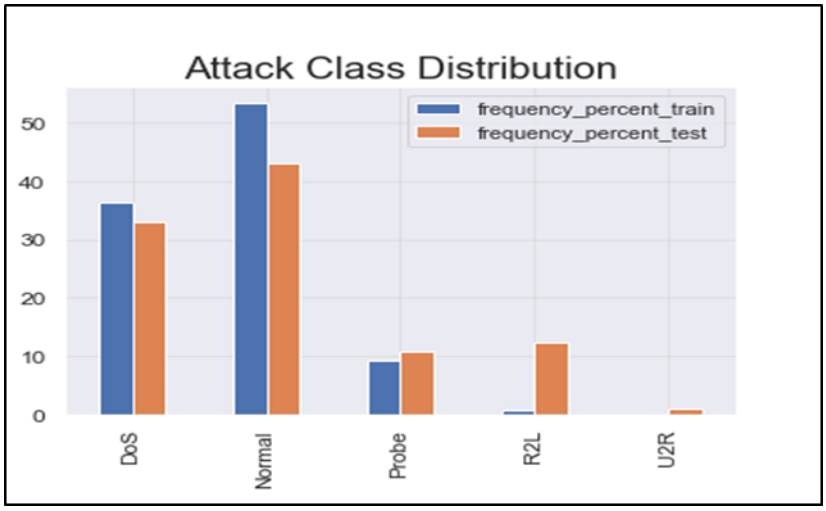

Fig 4: Attack Class Distribution

Table 4: Summarizes the performance metrics of all the classifier

S.No	Classifiers used	Accuracy in %	Precision in %	Recall in %	F1-Scorein %
1.	Naïve Bayes	42.90	.37	.43	.26
2.	Logistic Regression	75.18	.71	.75	.70
3.	SVM	75.48	.80	.75	.70
4.	KNN	75.32	.71	.75	.70
5.	Random Forest	73.35	.68	.73	.69
6.	XGBoost	70.77	.77	.71	.68
7.	C4.5	70.02	.64	.70	.66

As illustrated in Figure 5, the employment of the SVM classifier led to the highest level of accuracy. Figure 6 demonstrated that SVM surpassed other classifiers in terms of precision. Furthermore, our analysis from Figure 7 indicated that the SVM, KNN, and Logistic classifiers yielded the best results for recall.

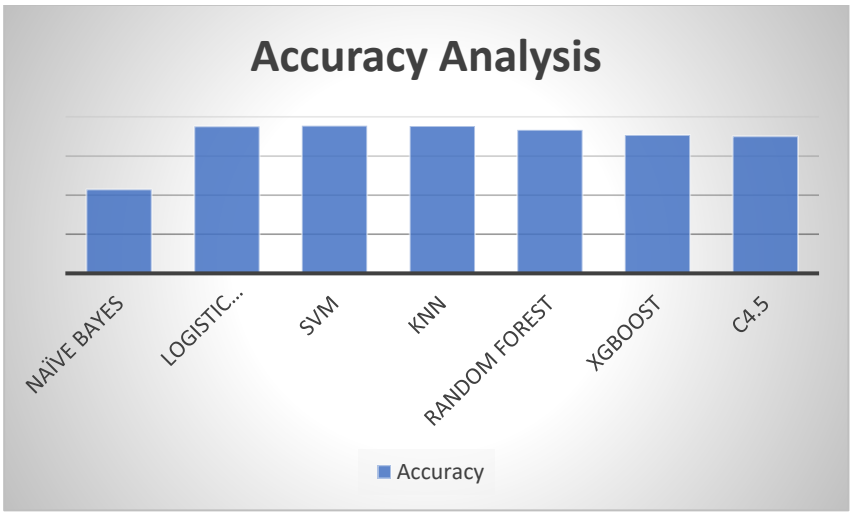

Fig 5: Accuracy Analysis

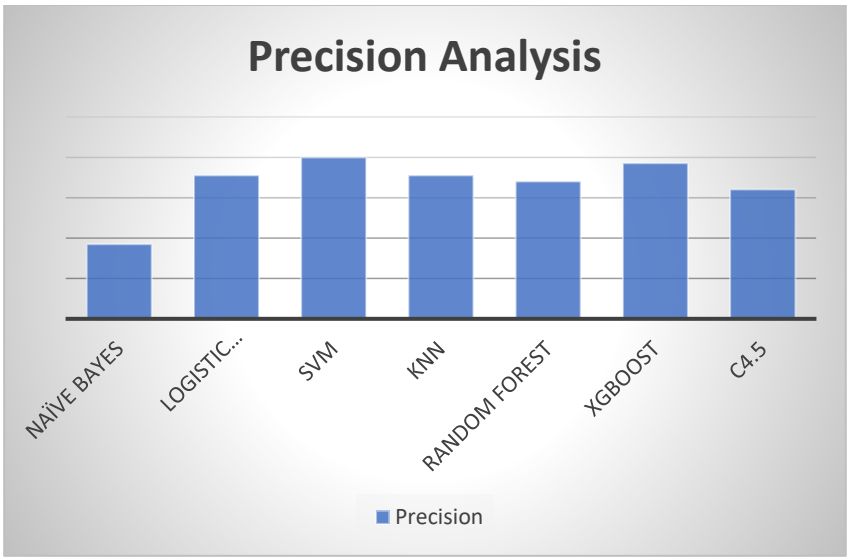

Fig 6: Precision Analysis

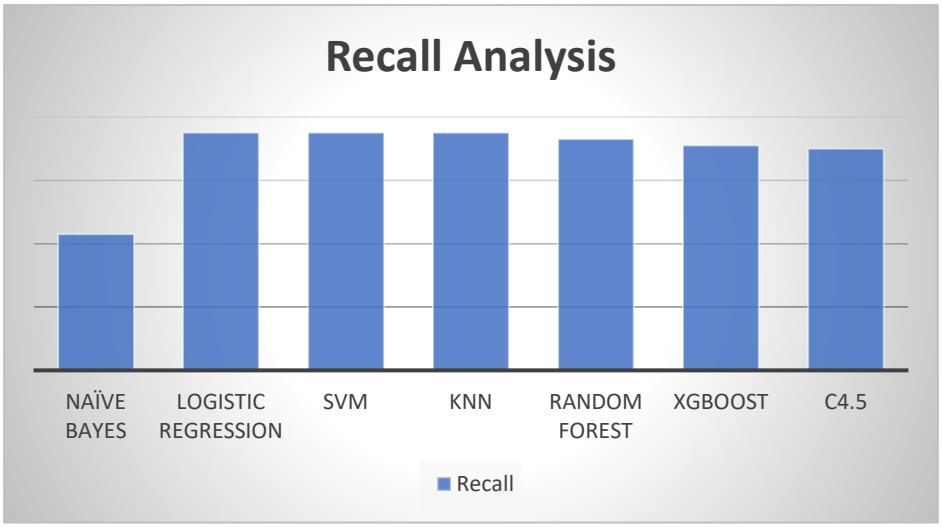

Fig 7: Recall Analysis

4. Conclusion and Future Work

We conducted a comparative analysis using Python's scikit module, implementing Naïve Bayes, Logistic Regression, SVM, Random Forest, C4.5, XG Boost, and KNN classifiers. After preprocessing the dataset, the SVM algorithm exhibited outstanding performance using NSL-KDD dataset, achieving the highest levels of accuracy,

precision, and recall. Although the KNN classifier closely followed the SVM in terms of accuracy, there was a significant difference in their execution times. Our experimental results lead us to the conclusion that SVM surpasses other classifiers in accuracy, precision, and recall. In our next plan for future, we plan to apply machine learning algorithms to real-time data and enhance their performance. Additionally, we aim to identify attacks not covered in the dataset.

References

[1] Muhammad Shafiq, Xiangzhan Yu, Asif Ali Laghari, Lu Yao, N abin Kumar Karn, F oudilAbdessamia, "Network Traffic Classification Techniques and Comparative Analysis Using Machine Learning Algorithms", 2016 2nd IEEE International Conference on Computer and Communications, vol. 8, pp. 2451-2455, 2016.

[2] JaiswalRupeshChandrakant, LokhandeShashikant. D., "Machine Learning Based Internet Traffic Recognition with Statistical Approach", 2013 Annual IEEE India Conference (INDICON), vol. 7, pp. 121-126, 2013.

[3] RiyadAlshammari, A. NurZincir-Heywood, "Identification of KDD encrypted traffic using a machine learning approach", Journal of King Saud University – Computer and Information Sciences, vol. 27, pp. 77–92, 2015.

[4] Alberto Dainotti, Antonio Pescapé, Kimberly C. Claffy," Issues and Future Directions in Traffic Classification", IEEE Network January/February 2012.

[5] T.NguyenandG.Armitage,"Asurveyoftechniquesf orInternettraffic classification using machine learning", IEEE Communications Surveys &Tutorials,Vol.10,No.4,fourthquarter2008,pp56-76.

[6] FatihErtam, Ilhan FiratKilinçer, Orhan Yaman,"Intrusion Detection in Computer Networks via Machine Learning Algorithms", International ArtificialIntelligenceandDataProcessingSymposium(IDAP),2017,pp 1-4

[7] Jamal H. Assi, Ahmed T. Sadiq, "NSL-KDD dataset Classification Using Five Classification Methods and Three Feature Selection Strategies", Journal of Advanced Computer Science and Technology Research, Vol.7 No.1, March 2017, 15-28.

[8] Muhammad Shafiq, Xiangzhan Yu, Asif Ali Laghari, Lu Yao, N abin Kumar Karn, FoudilAbdessamia, "Network Traffic Classification Techniques and Comparative Analysis Using Machine Learning Algorithms", 2nd IEEE International Conference on Computer and Communications,2016,pp2451-2455.

[9] Dewa, LeandrosMaglaras (2016) "Data Mining and Intrusion Detection Systems", International Journal of Advanced Computer Science and Applications, Vol 7 No 1,pp61-71.

[10] L. Dhanabal, and S. P. Shantharajah (2015) "A Study on NSL-KDD Dataset for Intrusion DetectionSystemBasedonClassificationAlgorithms" ,InternationalJournalofAdvanced Research in Computer and Communication Engineering, Vol 4, Issue 6,pp.

[11] Himadri Chauhan, Vipin Kumar, SumitPundir and Emmanuel S. Pilli (2013) "A Comparative Study of Classification Techniques for Intrusion Detection" International Symposium on Computational and Business Intelligencepp40-43.

[12] S. Revathi, Dr. A. Malathi (2013) "A Detailed Analysis on NSL-KDD Dataset Using Various Machine Learning Techniques for Intrusion Detection and Technology, IJERT Vol. 2 Issue 12 pp1848-1853.

[13] NSL-KDD dataset (Online Available): http://www.unb.ca/cic/datasets/nsl.html.

[14] Dhanabal, L., and S. P. Shantharajah. "A study on NSL-KDD dataset for intrusion detec- tion system based on classification algorithms." International Journal of Advanced Re- search in Computer and Communication Engineering 4.6 (2015):446-452.

[15] Revathi, S., and A. Malathi. "A detailed analysis on NSL-KDD dataset using various ma- chine learning techniques for intrusion detection." International Journal of Engineering Re- search and Technology. ESRSA Publications(2013).

[16] Narayan, Vipul, et al. "A Comprehensive Review of Various Approach for Medical Image Segmentation and Disease Prediction.

[17] Mall, Pawan Kumar, et al. "A comprehensive review of deep neural networks for medical image processing: Recent developments and future opportunities." Healthcare Analytics (2023): 100216.

[18] Narayan, Vipul, et al. "Severity of Lumpy Disease detection based on Deep Learning Technique." 2023 International Conference on Disruptive Technologies (ICDT). IEEE, 2023.

[19] Saxena, Aditya, et al. "Comparative Analysis Of AI Regression And Classification Models For Predicting House Damages İn Nepal: Proposed Architectures And Techniques." Journal of Pharmaceutical Negative Results (2022): 6203-6215.

[20] Kumar, Vaibhav, et al. "A Machine Learning Approach For Predicting Onset And

Progression""Towards Early Detection Of Chronic Diseases "." Journal of Pharmaceutical Negative Results (2022): 6195-6202.

[21] Chaturvedi, Pooja, A. K. Daniel, and Vipul Narayan. "A Novel Heuristic for Maximizing Lifetime of Target Coverage in Wireless Sensor Networks." Advanced Wireless Communication and Sensor Networks. Chapman and Hall/CRC 227-242.

[22] Kumar, Vimal, and Rakesh Kumar. "A cooperative black hole node detection and mitigation approach for MANETs." In Innovative Security Solutions for Information Technology and Communications: 8th International Conference, SECITC 2015, Bucharest, Romania, June 11-12, 2015. Revised Selected Papers 8, pp. 171-183. Springer International Publishing, 2015.

[23] Kumar, V., Shankar, M., Tripathi, A.M., Yadav, V., Rai, A.K., Khan, U. and Rahul, M., 2022. Prevention of Blackhole Attack in MANET using Certificateless Signature Scheme. Journal of Scientific & Industrial Research, 81(10), pp.1061-1072.

[24] Kumar, V. and Kumar, R., 2015. An adaptive approach for detection of blackhole attack in mobile ad hoc network. Procedia Computer Science, 48, pp.472-479.

[25] Kumar, V. and Kumar, R., 2015, April. Detection of phishing attack using visual cryptography in ad hoc network. In 2015 International Conference on Communications and Signal Processing (ICCSP) (pp. 1021-1025). IEEE.

[26] Kumar, V. and Kumar, R., 2015. An optimal authentication protocol using certificateless ID-based signature in MANET. In Security in Computing and Communications: Third International Symposium, SSCC 2015, Kochi, India, August 10-13, 2015. Proceedings 3 (pp. 110-121). Springer International Publishing.

[27] Kumar, V. and Kumar, R., 2017. Prevention of blackhole attack using certificateless signature (CLS) scheme in MANET. In Security Solutions for Hyperconnectivity and the Internet of Things (pp. 130-150). IGI Global.

[28] Gupta, P., Kumar, V. and Yadav, V., 2021. Student's Perception towards Mobile learning using Interned Enabled Mobile devices during COVID-19. EAI Endorsed Transactions on Industrial Networks and Intelligent Systems, 8(29), pp.e1-e1.

[29] Deshwal, V., Kumar, V., Shukla, R. and Yadav, V., 2022. Estimating COVID-19 Cases Using Machine Learning Regression Algorithms. Recent Advances in Electrical & Electronic Engineering (Formerly Recent Patents on Electrical & Electronic Engineering), 15(5), pp.390-400.

[30] Mr. Kaustubh Patil. (2013). Optimization of Classified Satellite Images using DWT and Fuzzy Logic. International Journal of New Practices in Management and Engineering, 2(02), 08 - 12. Retrieved from http://ijnpme.org/index.php/IJNPME/article/view/15

[31] Pathak, D. G. ., Angurala, D. M. ., & Bala, D. M. . (2020). Nervous System Based Gliomas Detection Based on Deep Learning Architecture in Segmentation. Research Journal of Computer Systems and Engineering, 1(2), 01:06. Retrieved from https://technicaljournals.org/RJCSE/index.php/journal/article/view/3