SEMARAK ILMU
PUBLISHING
202103268166(003316878-P)

Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

Big Data Analysis on Network Intrusion Detection using High Performance Deep Neural Networks

Rajendran Bhojan[1,*], Saravanan Venkataraman[2]

[1] Department of Mathematics & Computer Science, Papua New Guinea University of Technology, Lae 411, Papua New Guinea
[2] College of Technology and Business, Riyadh ELM University, Qurtubah, Riyadh 13244, Kingdom of Saudi Arabia

ABSTRACT

The rapid evolution of the internet over decades has given rise to a significant increase in cyber-attacks, propelled by the growth of high-speed internet. This paper addresses the escalating threat by focusing on Network Intrusion Detection Techniques in big data environments, specifically utilizing high-performance deep neural networks for analysing intrusive data within the enhanced NSL-KDD dataset. The research emphasizes the application of the random forest algorithm to enhance accuracy in detecting intrusive attack data within large datasets. The proposed model is thoroughly evaluated using the enhanced NSL-KDD dataset, demonstrating superior performance compared to existing systems. The study contributes a comprehensive overview of the methodology, incorporating advanced techniques in deep neural networks and random forest algorithms. The results highlight the effectiveness of the proposed model in bolstering cybersecurity measures, offering a robust solution for the evolving landscape of cyber threats.

Keywords:
Deep neutral network; Intrusion detection; NSL-KDD dataset; Random forest

1. Introduction

Global infrastructure under an extreme critical part that has become a network, based on the facts of business data, economic, personal and banking in the computer networks has been shared with the major aspect of internet under security. NID (Network Intrusion Detection) in network security has become a crucial catalyst. With network security, communities which face challenges of such attacks, create awareness to mitigate and also the prevent attackers. They discover new vulnerabilities every year in compounding numbers. The new attacks in turn are meant for security tools to detect through automation process. From such attacks, computer networks have become very significant with the role of intrusion detection systems. From the public networks, in order to protect data, network security is a defending measure in the use of firewalls in all companies and organisations around the world. True security of 100 per cent cannot be obtained under the resources of security from the user to secure the resources that can be used through firewall. From

[*] *Corresponding author.*
E-mail address: rajendran.bhojan@gmail.com

https://doi.org/10.37934/araset.55.1.167176

the activity of normal signatures in various signature activities of attacks based on the supposition is worked with the tool of NID. For detecting attacks, two ways of means may be Network Intrusion Detection or anomaly-based detection. To determine the attack matches to detect against a database of recognition of signature attacks are used in the analysis of signature based on abnormal behaviour issue and a normal baseline against the monitoring are used to detect the anomaly.

The detections on attacks consist of two main challenges. First, rapid production and large traffic network within different domain that are used in wide technologies under the analysis in efficient deep learning techniques, data analysis in these networks are of very large and high volumes in processing. Second, in various scenarios and pattern in different types of attacks is the problem of feature selection, to intrusion detection applied under machine learning in many methods, like ANN (Artificial Neural Network), RF (Random Forest), SVM (Support Vector Machine) and so on is detection of accuracy to improve the classification techniques. Similarly, with better performances classification, the unrelated data are eliminated in the feature selection while selecting key attributes and easy understanding of the related data in good way, rather than processing unnecessary data. In performances, generalization has a vital impact on quality of features, to make mistakes prone to design, good features with uneasy way from unlabelled data under a great number in representing the better features is the designing approach of deep neutral network. The study addresses the escalating cyber threats in the era of internet growth by emphasizing the need for robust Network Intrusion Detection Techniques (NIDTs). In the context of high-speed internet and big data environments, existing systems struggle to accurately detect intrusive attacks. The primary objective is to propose an advanced Intrusion Detection System (IDS) model, utilizing the Random Forest algorithm for effective classification within the NSL-KDD dataset. The study aims to enhance accuracy, precision, recall, and F1-score in detecting various attack categories, contributing to the evolution of resilient Network Intrusion Detection Techniques.

This paper for NID system proposes high performance of deep neutral network, for the undefined previous attacks based on the adjusted and learned features. In particular, the key features under the extract and the input data, to learn the features with the random forest algorithm in encoding process using enhanced NSL-KDD dataset of intrusion detection evaluated to propose the model usability.

2. Literature Survey

Nour Moustafa *et al.,* [1], this paper is based on intrusion detection system with big data analysis using finite Dirichlet mixture models. To detect the cyber domain of malicious activities that has become a vital mechanism as an intrusion detection system. For the attacks detected the outputs come as an important limitation in the system with relatively high false alarm rate that reduced the relative. Network data analysis and the task monitoring are no longer considered under the necessary isolation, for identification of event anomaly with methods for decision making under their integrated optimization. Based on the technicians of anomaly detection under the Dirichlet mixture model with a new engine of statistical decision, pre-processing, logging and capturing were the three modules of framework. The statistical analysis of data network that represent the result under the empirical that helps in selecting the best model for the network data. Based on the correlation under these techniques and those cannot detect the distance measures under model attacks with normal activities, using the model of mixed Dirichlet and fro finding small various under the range of boundaries that identify in utilising the attacks.

S.S. Panwar *et al.,* [2], this paper is based on the reduction techniques of data from the NSL-KDD dataset analysis. The amount of available data is huge under the information field that need to be turned as a useful information. Techniques and data reduction has been used in this process. Data reduction is known as the minimizing and maximizing the amount of data process to store in a storage data environment. In computer network they have been achieved under different data reduction or process with cost of computational reduction and increasing in efficient storage. On NSL-KDD dataset the algorithm of data reduction that has been applied. Algorithm of data reduction techniques for data reduction that enhances has been useful of the classified algorithm under performances. Comparisons are based on the result under the sensitivity, specificity and accuracy.

Suad Mohammed Othman *et al.,* [3], this paper is based on big data environment using machine learning algorithm model intrusion detection. For big data that have changed the system, data analysis and security information in the system or network to detect any intrusion under data analysis and monitor the system of IDS (Intrusion Detection System). By traditional techniques to detect attacks under the process of data analysis that have made in the network with data generation as high speed and high volume. Techniques of big data are used in IDS to deal with process of efficient data analysis and accuracy of big data. By using SVM in the model intrusion detection are built and for feature selection under the model were used. For big data it is efficient and training time is reduced, as the high performances model spark-chi-SVM that represent the experiment.

M. Mazhar Rathore *et al.,* [4], this paper is based on the big data environment for ultra-high speed with real time intrusion detection system. In increasing, day to day as network services and internet with number of people, very high speed over the internet networks those are generated, over the large amount of data. The enterprise network, website, and internet threats on the internet, with more security in real time, task challenging in ultra-high-speed environment under detection intrusion. Using machine learning approach for different types of networks was proposed. To overcome the challenges that do not provide real time detection, recent unknown attacks are unable to detect. Using Hadoop Implementation for ultra-high speed big data environment real time intrusion detection, from the analysis of DARPA datasets using BER and FSR for classification, under the select nine parameters was proposed in the feature selection scheme. In terms of accuracy the best classifiers under all these classifiers are among the result. With traditional techniques, by comparing result in terms of processing time and efficiency with respect to the terms of accuracy as evaluated was proposed in the architecture system.

Mahzad Mahdavishariff *et al.,* [5], this paper is based on communication network under the intrusion detection system in big data as deep learning approach. In defence system they used computer networks and databases under the computer network status about the information that are considered that hackers under most important parameter the security perspective from possible intrusive patterns to mine in this growing mountain of data, under the mechanism of intrusion detection. Using big data techniques to design a system of intrusion detection that is necessary to design under the big data nature can be handled with bid data techniques to develop an intrusion detection system. To cope with these challenges, IDS (Intrusion Detection System) for effective and efficient design of deep learning methods, big data awareness was employed. The system of intrusion detection designed to increase in accuracy and false alarm that could be reduced. The speed of algorithm in deep learning that can be improved, in techniques of big data analysis, due to high complexity that have low execution speed.

Tongtong Su *et al.,* [6], this paper is based on the NSL-KDD dataset on Network intrusion detection under deep learning methods. Network security that has been an effective and network traffic under the unknown attacks can be identified with intrusion detection. Based on models of traditional machine learning for detection under existing methods the design of traffic features under relies heavily and low accuracy, they get a relatively good performance that can obtain some outstanding features. In intrusion detection as feature engineering and low accuracy as to solve a problem, this paper proposed in BAT model a detection traffic anomaly. For classification network traffic that can be obtained key features by the BLSTM model that is generated under packet vector that composed the network flow vector that used in attention mechanism. For network traffic classification they used the SoftMax classifier.

Sameel Al *et al.,* [7], this paper is based on the big data environment for imbalanced data under the intrusion detection of system under a new hybrid network. In recent years the intrusion detection system and big data analysis has been used under the deep learning approach. On network flow traffic generating big data which is proposed in the detection system based on network attack under new classification. For a good intrusion system LSTM and CNN has been considered under the hybrid deep learning proposed system. The performances of the system with imbalanced data to reduce the effect that has been used under the sampling method that is used in SMOTE considers in the process of imbalanced data. The proposed method has been compared with the algorithm of deep learning and machine learning.

Xiaoming Li *et al.,* [8], this paper is based on deep learning in digital twins of the internet o things under big data analysis. In smart cities on massive data generation under BDA (Big Data Analysis) with IoT, safe data processing and efficiency to the time of governances of direction, in smart city changes. In smart cities under the multi-source collection of data using BDA, the algorithm of deep learning ad CNN (Convolutional Neutral Network) strategy, through forward distribution was introduced. Based on DL multi-hop transmission of IoT-BDA system, to construct the smart city as technology if multi-hop transmission. The system performances were analysed and simulated. Data transmission model in the energy efficiency reveals the result in increasing the energy efficiency and decreasing the energy. The IoT-BDA system to signal energy efficiency transmission under the power diversion factor was more suitable. The constructed system with regard to performance of data transmission is to adopting the DL algorithm by other scholars. Using DL approach, the delay in reducing the data transmission is to improve the smart city system Internet of things –BDA.

Fiona X. Yang *et al.,* [9], his paper is based on the AI-based big data analysis for customers making process under beauty premiums. The process of decision making in various stages is by placing facial features under different aesthetics. System of facial recognition under AI (artificial Intelligence) is to facilitate facial incorporations in developing a comprehensive model. On purchase they have a positive effect to have smile and beauty scores that represented the result and to service cures subject the post service rating. The context in AI-based facial analysis in pioneering study in research and the impression management offers insights.

Yousef Methkal Abd Algani *et al.,* [10], this paper is based on the big data analytics under anomalous behaviour of wireless networks. The global throughput under extensive cellular technologies and internet connection is assessed. In the cyber space domain from detection of a range of broad hostile activities is an essential tool for anomaly detection. Crucial inputs increase under the ADS (Anomaly Detection System) that reported every data vulnerable and defective cyber security. The result of anomalous assaults that could be detecting unusual behaviour and activity scanning network that is developed, is the major objective here. With great precision for each assault to recognize with an appropriate mixture in selecting strategies in several characteristics, combine the method of anomaly detection. Using NSL-KDD dataset, the effectiveness of method suggested is assessed. In all sorts of attacks in minimal percentages of false positive and by excellent retaining precision, has proved the ADS effectiveness. Research Gap.

Despite the abundance of studies addressing intrusion detection systems (IDS) in big data environments, a noticeable research gap exists in achieving a balance between high accuracy and low false alarm rates. Current approaches often struggle with imbalanced datasets and fail to adequately address the challenges posed by unknown or real-time attacks. Additionally, there is a need for more comprehensive evaluations of hybrid models that integrate machine learning algorithms with traditional techniques to enhance the overall effectiveness of intrusion detection systems.

3. NSL-KDD Dataset

The proposed IDS model is tested with NSL-KDD dataset to get trained. The NSL-KDD is an updated version of KDD cup 99 dataset for intrusion detection system of standard NSL-KDD by McHugh to solve the discussed problem, of KDD cup 1999 based on problem [11]. A small portion of dataset is selected randomly to run the experiment. The entries contain 4,898,531 datasets of the NSL-KDD dataset. Independent of operating systems or applications, raw network packets are collected from dataset of NSL-KDD. Consequently, in this dataset of each record, a label of each class to be recognised has been provided. To be precise in the dataset in all labels, 37 types of attacks contained in NSL-KDD. Four categories precisely fell in simulation. The attack categories were: remote to local, user to root, probing attack, and denial of services.

During the training process which could not be represented with 41 features contained in the original NSL-KDD dataset of each record, the data pre-process and symbolic features of three extended flag with 1-N encoding, services with symbolic features of protocol types. With a max-mix operations to the range [0,1] the dataset of NSL-KDD is normalized.

Table 1
NSL-KDD attacks

Attacks in Dataset	Types of attacks
DoS	Worm, Udpstor m, Processtable, Mail bomb, Pod, Smurf, Neptune, Land, Back, Apache2, Teardrop
Probe	Satan, Nmap, Portsweep, Mscan, Sa int, IPsweep
R2L	Guess_password, Warezmaster, Multihop, phf, Ftp_write
U2R	Buffer overflow, Xterms, Ps, Loadmodule Rootkit, Perl, Sqlattack

4. Proposed Model

4.1 Random Forest Algorithm

This is a Supervised Classification algorithm. By constructing data to be classified, the random forest algorithm achieves higher level of prediction [12]. The class label in order to predict is combined with the resultant individual, which are constructed as trees with the test dataset by applying random forest techniques [13]. It is not worthy with a single classifier to classify huge volumes and amounts of data which might result in less accuracy. With the algorithm of decision tree classification and compounding with huge volumes of data in any application [14], the random forest classifier algorithm constructs the decision tree [15].

i. In the samples of data instances, number of training data instances is taken as N. This is given as input dataset with number of attributes M.
ii. Each tree nodes are chosen as the next set of attributes which are determined inputs, as the number of parameters m.
iii. A tree is constructed by replacement of each sample and taken along with training sample.
iv. For a free node, in that particular node *m* attributes are arbitrarily selected.
v. Best split is calculated based on *m* input attributes with the sample dataset.
vi. Without pruning it is grown in each tree.

The supervised classification algorithm is one of the random forest algorithms. The proposed model is to enhance the generalization ability in joining their outputs to building several trees that are used in decision tree algorithm as the random forest. To produce a strong learner tree under several individual trees by combining the random forest algorithm. To obtain one stranger learner tree under several classification trees generating the random forest algorithm, by using a tree classification algorithm from the original dataset using a different bootstrap sample by each tree in order to construct working random forest. The purpose of classification process in the forest at every tree is placed down to be as classifiers that require a new object in the construction of forest. To develop IDS system that has been used in the random forest algorithm.

```
Algorithm: Random Forest Algorithm
Input: Training data T, Parameter ($\delta, k, s, c, \lambda$)
Output: Model with evaluation
Step 1: Ensemble-RF (T, $\delta, k, s, c, \lambda$)
Step 2: for 1< -1 to s do
Step 3:     (train, test) ← randomSplit (T, $\lambda$)
Step 4:     split ← bootstrap (train,
Step 5:     model ← Random Forest. train (split, c)
Step 6:     score ← evaluate (model, test)
Step 7:     out [i] ← (model, score)
Step 8:  end for
Step 9: return out
```

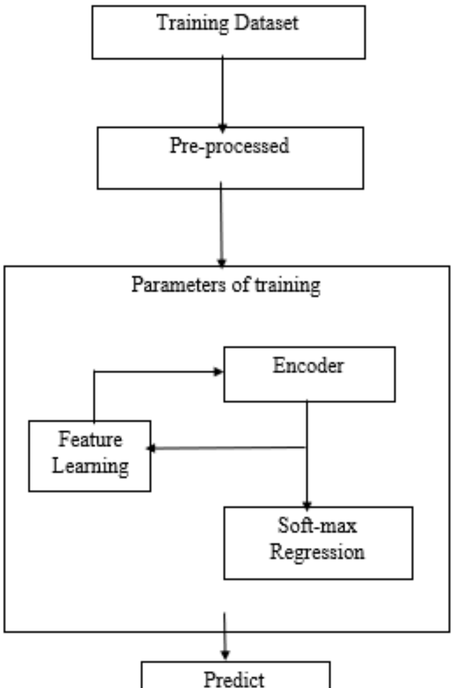

Fig. 1. Model Classification

The data to be trained in first pre-processor as represented in the model is the data input to process inside the function. Among various features of the complex relationship that encoded the learning, key features have been extracted and by a soft mac regression under the prediction and final classifier are resulted [12].

The model is constructed as one input layer, hidden layers and one output layer. Accordingly, ideas of compressing features, five features are reduced as the input of 122 features. At the end to reduce the five learning features until in the sequence of operations in performing the hidden layers remaining and layer i, the soft-max of hidden 5 layer, to the soft-max layer as inputs that are used with these features, the input data under classifier predictions under the soft-max layer and the results in getting final output layer [13]. A complete network model is implemented as network stacks in these as hierarchical.

5. Result Explanation

5.1 Performances Metrics

The results of the proposed model to test have been carried out under the performance measures such as time, true positive, false Positive and accuracy [14]. The performance measures under the equation as follows:

Accuracy: Precise classification under measure of proportion. The sentiment classification helps in the measure of accuracy. For overall process, the model with accurate ratio is used for estimation. The accuracy is defined as calculation below:

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \tag{1}$$

Precision: by the model proportion that identified the sample of true positive samples. It is calculated on the class that depending on the negative or positive based on the label under the value of prediction estimate the precision.

$$\text{Precision} = \frac{TP}{TP+FP} \tag{2}$$

Recall: by the model as positive sample that are identified the ratio of positive sample. The corrected classification based on the tweets ratio as mentioned in the recall. By using the formula that measure as recall as follows,

$$Recall = {}^{TP}/_{TP} + FN \tag{3}$$

F1-Score (F1): The performances that evaluated the use in it. To calculate the integrate it is defined with F1-score, recall and precision as follows,

$$F1 \ = 2 \ \times \frac{precision \times recall}{precision + recall} \tag{4}$$

True Negative (TN): normal record as valid records under correct classifier
True Positive (TP): attacks as attacks records under correct classifier
False Positive (FP): attacks as normal data records under the incorrect.
False negative (FN): normal record as record attack under the incorrect.

The model that can be evaluated as achievability with accuracy, but it is trained with sample data when it is unbalanced, the evaluation of the model will not be done properly as it is necessary to add recall and precision in practical applications that have various emphasis under high recall and high accuracy. Precision and recall were the mean harmonic in F1-score. The model has the ability in generalizing the reflection [15].

5.2 Performance Evaluation

To evaluate the performances of classification model under the experiment within 5 class categories, the sample 22544 model is tested under the classification and the experimental result according to the generated confusion matrix is represented in the Figure 2. The normal sample 9711 from dataset were correctly identified under the normal sample 9433, while identified correctly with the 12833 attacks under the remaining 8554.

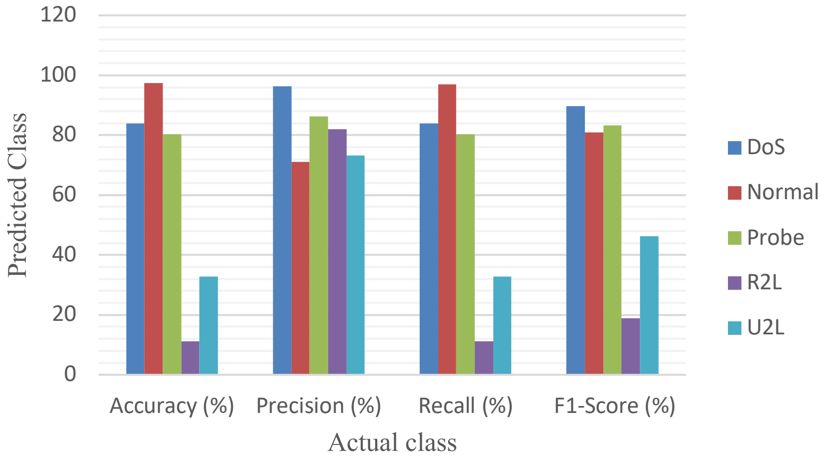

Fig. 2. Performances of the classes

According to the Eq. (1) to Eq. (4), it can be obtained from Accuracy, Precision, Recall and F1score as explained in the Table 2. Since the sample available has been limited for training within five classes, low attacks were under the rate of detection of R2L and U2R, accuracy of NIDS has been reduced.

Table 2
Performances of 5 class categories

Category	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
DoS	83.9	96.33	83.95	89.70
Normal	97.30	71.01	97.03	80.98
Probe	80.37	86.37	80.37	83.27
R2L	11.26	82.03	11.28	18.79
U2L	32.80	73.32	32.83	46.37
Total	79.73	82.23	79.73	76.46

Other similar methods are compared with the deep learning based in order to ensure the effectiveness in our model using NSL-KDD dataset in the 5-class classification under the F1-score as they produced 75.76%. They are not given under the results of precision and recall by their graph as represented as they are of 84% and 68% respectively. The precision of 83%, under the achiever’s model is better than the recall and f1-score of 80% and 77% respectively.

Based on the dataset in the 5 class performances with the accuracy of 76% achieved in the method DNN. This method has achieved 80% of the lower result.

6. Conclusion

The network intrusion behaviour for detection has been proposed for NID approach in high performances deep neutral network. For patterns undefined previously, this method adjusts itself and to learn networks used in this method. With the method proposed that can be reduced false negative and alarm the possibility. To construct a network, benefits of random forest taking encoder predictions with simple network structure and a large flow of network model exploration. For improvements efforts will be made in detecting the higher degree of accuracy with the existing model with other classifiers by detecting the real time problem.

Acknowledgement

This research was not funded by any grant.

References

[1] Moustafa, Nour, Gideon Creech, and Jill Slay. "Big data analytics for intrusion detection system: Statistical decision-making using finite dirichlet mixture models." *Data Analytics and Decision Support for Cybersecurity: Trends, Methodologies and Applications* (2017): 127-156. https://doi.org/10.1007/978-3-319-59439-2_5

[2] Panwar, Shailesh Singh, and Y. P. Raiwani. "Data reduction techniques to analyze NSL-KDD Dataset." *Int. J. Comput. Eng. Technol* 5, no. 10 (2014): 21-31.

[3] Othman, Suad Mohammed, Fadl Mutaher Ba-Alwi, Nabeel T. Alsohybe, and Amal Y. Al-Hashida. "Intrusion detection model using machine learning algorithm on Big Data environment." *Journal of big data* 5, no. 1 (2018): 1-12. https://doi.org/10.1186/s40537-018-0145-4

[4] Rathore, M. Mazhar, Awais Ahmad, and Anand Paul. "Real time intrusion detection system for ultra-high-speed big data environments." *The Journal of Supercomputing* 72 (2016): 3489-3510. https://doi.org/10.1007/s11227-015-1615-5

[5] Mahdavisharif, Mahzad, Shahram Jamali, and Reza Fotohi. "Big data-aware intrusion detection system in communication networks: a deep learning approach." *Journal of Grid Computing* 19, no. 4 (2021): 46. https://doi.org/10.1007/s10723-021-09581-z

[6] Su, Tongtong, Huazhi Sun, Jinqi Zhu, Sheng Wang, and Yabo Li. "BAT: Deep learning methods on network intrusion detection using NSL-KDD dataset." *IEEE Access* 8 (2020): 29575-29585. https://doi.org/10.1109/ACCESS.2020.2972627

[7] Al, Samed, and Murat Dener. "STL-HDL: A new hybrid network intrusion detection system for imbalanced dataset on big data environment." *Computers & Security* 110 (2021): 102435. https://doi.org/10.1016/j.cose.2021.102435

[8] Li, Xiaoming, Hao Liu, Weixi Wang, Ye Zheng, Haibin Lv, and Zhihan Lv. "Big data analysis of the internet of things in the digital twins of smart city based on deep learning." *Future Generation Computer Systems* 128 (2022): 167-177. https://doi.org/10.1016/j.future.2021.10.006

[9] Yang, Fiona X., Ying Li, Xiaotong Li, and Jia Yuan. "The beauty premium of tour guides in the customer decision-making process: An AI-based big data analysis." *Tourism Management* 93 (2022): 104575. https://doi.org/10.1016/j.tourman.2022.104575

[10] Abd Algani, Yousef Methkal, G. Arul Freeda Vinodhini, K. Ruth Isabels, Chamandeep Kaur, Mark Treve, B. Kiran Bala, S. Balaji, and G. Usha Devi. "Analyze the anomalous behavior of wireless networking using the big data analytics." *Measurement: Sensors* 23 (2022): 100407. https://doi.org/10.1016/j.measen.2022.100407

[11] Tamilarasi, K., K. Maheswari, S. Ramesh, Samson Isaac, and A. Rajaram. "A Decentralized Smart Healthcare Monitoring System using Deep Federated Learning Technique for IoMT." (2023).

[12] Jaafar, Nurulaini, Siti Rohani Mohd Nor, Siti Mariam Norrulashikin, Nur Arina Bazilah Kamisan, and Ahmad Qushairi Mohamad. "Increase students' understanding of mathematics learning using the technology-based learning." *International Journal of Advanced Research in Future Ready Learning and Education* 28, no. 1 (2022): 24-29.

[13] Adnan, Ahmed Yaseen, Mustaffa Kamal Iwan, and Hannan Mohammed Abdul. "Intelligent control for ship manoeuvering." *Journal of Advanced Research in Applied Mechanics* 67, no. 1 (2020): 1-9. https://doi.org/10.37934/aram.67.1.19

[14] Yusof, Ahmad Anas, Saiful Akmal Sabaruddin, Syarizal Bakri, and Suhaimi Misha. "Simulation of System Pressure Impact on the Water Hydraulic Hybrid Driveline Performance." *CFD Letters* 10, no. 2 (2018): 59-75.

[15] Yagoub, Sami Abdelrahman Musa, Gregorius Eldwin Pradipta, and Ebrahim Mohammed Yahya. "Prediction of bubble point pressure for Sudan crude oil using Artificial Neural Network (ANN) technique." *Progress in Energy and Environment* (2021): 31-39.