Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.98549909

A MODEL-DRIVEN FRAMEWORK FOR DARKNET NETWORK PREDICTION USING NEURAL NETWORK

M. S.Bennet Praba[1],*M.S.Antony Vigil[2], and W.Ancy Breen[3]
[1,2,3]Department of Computer Science and Engineering,
[1,2, 3]SRM Institute of Science and Technology, Ramapuram, Chennai, Tamilnadu.
*Corresponding Author: M.S.Antony Vigil, Email id: antonyvigil@gmail.com

Abstract:
Internet of Things (IoT) devices is vulnerable to assaults such as Darknet and blackhole attacks because of their restricted capabilities. Using the CIC-Darknet dataset, DarkWeb Traffic Detection System (DTDS) models are developed and evaluated using machine learning and deep learning techniques. Using DTDS models, we were able to identify and categories Darknet activity in IoT networks. Machine learning helps keep sensitive information safe and enhances network performance. When applied to Darknet data, DTDS models improve IoT security by identifying and classifying threats. IoT devices are vulnerable to cyberattacks like Darknet or blackhole attacks, is explored, and DTDS in IoTs effectiveness is assessed. The research concludes that DTDS models may be used to effectively detect and classify darknet traffic in IoT networks, leading to improved network services and data security. The performance of DTDS models was measured using accuracy, precision, recall, and F1-score, and the models were shown to be superior to their rivals.
Keywords: Internet of Things, Machine Learning, Dark Web Traffic Detection System, Darknet

INTRODUCTION:

Internet of-Things (IoT) advances can possibly improve personal satisfaction [1] because of their capacity to assemble and analyze information about the general climate. This works with the improvement of "brilliant urban areas" by making it more straightforward for individuals and things to speak with each other. The quantity of associated gadgets is supposed to arrive at 50 billion by 2020 [2, 3]. The IoT is an intricate interconnected network. Securing an IoT system might be difficult due to the large attack surface. Due to its widespread use, deploying IoT solutions has evolved into a unified procedure. Security, energy efficiency, analytics methods, and compatibility with other software applications are just some of the many considerations that need to be made when establishing an IoT system [3].

The gadgets that make up the Internet of Things (IoT) are becoming autonomous. Therefore, a thief might potentially get access to these tools. Devices in the Internet of Things often communicate via some kind of communication channel, which may make it easy for hackers to eavesdrop and steal private information. Finding a reference design that can handle both existing features and improvements to those features will be challenging as the IoT industry develops. Therefore, such an architecture needs to be distributive so that an environment can be built in which information are handled by different elements in a conveyed way in the wake of being gotten from different sources, versatile so a rising number of gadgets and

administrations can be upheld without debasing execution, interoperable so gadgets from various producers can cooperate to accomplish shared objectives, and asset effective so that it can run with the minimum amount of resources possible. There have been many attempts to standardise things, but so far there is no agreed-upon reference architecture, and developing one has been difficult. The fundamental issue is the inescapable dispersion of possible applications, each of which relies on a diverse set of components and design norms. The marketing strategy of any service provider must be connected to this problem [3].

A significant number of cyberattacks are being launched against IoT devices because they may be tricked into sending spam emails. As long as it can connect to the internet, a device may be infected and used to create a botnet. Because of their poor security, Internet of Things gadgets are often exploited as bots. Many dark market sites host botnets and other malicious software. Source code for botnets, such as the one used in the Mirai botnet, may be purchased or even made public. The cost of launching a DDOS attack may vary from the tens to the many dollars [7,8,9] relying upon the sort of administration, the quantity of bots or gadgets available for use in the botnet, and the seriousness and length of the assault. Some botnets even compete with one another on the Darknet due to the high level of competition there. After a successful assault on an IoT device, another botnet may try to "fix" the security hole that the successful botnet exploited to prevent future infections and keep control of the device by replacing the original infection with its own malware. The Dark Web, also known as the Deep Web or the Darknet, is a network of websites that anybody may visit, but whose IP addresses are concealed. The deep web is expected to dwarf the shallow web in size [10]. As the Internet has rapidly evolved since then, this figure has only risen.

2.RELATED WORK:

Large-scale deep learning networks for marine ship object detection in radar pictures are difficult to deploy on presently available radar technology. Incorporating thick associations, remaining associations, bunch convolution, stem blocks, and extractor modules, the LiraNet network is proposed in this exploration. The proposed two-way thick association module is carried out in the extractor organization, in this manner diminishing the functional intricacy of the framework. The created stem block utilizes a progression of little convolutions to remove the qualities of the information picture. This study proposes Lira-you just look once (Lira-Consequences be damned), a lightweight modelthat combines LiraNet with the object identification framework Darknet to facilitate ship recognition in radar images [11].

A new kind of proactive defence has arisen, and it's called moving target defence (MTD). The main idea behind MTD is to make attacks more unpredictable and confusing for aggressors by changing the assault surface (i.e., framework or organization arrangements), which could deliver the data assailants have assembled pointless and keep assaults from being done, prompting the disappointment of the assault. Ongoing developments in programming characterized organizing (SDN) innovation, especially regarding programmability and controllability with the help of SDN regulators, have made an assortment of mind-boggling framework operations highly flexible and dependable. As a result, several security operations have taken advantage of the SDN features' ideal deployment in a complex network. In this research, we provide an MTD approach that reconfigures a host's organization boundaries (like

its Macintosh, IP, and port numbers) contingent upon its criticality, utilizing an assault chart. At the point when the host is on the assault path(s), this MTD approach based on the attack graph is particularly vulnerable to assault.[12]

Academics are interested in it as one of the key research topics for the ICN due to its efficient forwarding mechanism. NDN is vulnerable to cyberattacks in the same way that any other network [18]. With the aid of the cooperating server, a CIFA may exploit vulnerabilities in the NDN's interior sending component to communicate noxious interest parcels as a heartbeat. This makes veritable clients' normal solicitations be blocked and diminishes the nature of administrations given by the NDN organization. By inspecting the qualities of the organization traffic and the CIFA model, a clever technique for distinguishing CIFA is made, in view of the expectation blunder between a molecule channel and a one-step expectation calculation.[3]

The Internet of Things relies heavily on wireless sensor networks, yet these networks are vulnerable to cyberattacks and have poor communication routes. Motivated by this problem, a set of robust routing strategies for WSNs is proposed. The core idea is to use connection reliability with more typical routing measures like distance and delay when developing routing algorithms. To begin, a new profound learning-based interface forecast model was proposed, which joins the Weisfeiler-Lehman portion and Double Convolutional Brain Organization (WL-DCNN)to extract and identify small, lightweight subgraphs. It's utilised to make the topological feature mining technique more capable of learning on its own and being used more broadly. Experimental results on six different open complex networks datasets reveal that WLDCNN outperforms nine baseline techniques [14].

In earlier work, an essential element of the remote organization situation is the presence of a heterogeneous remote organization region, which results from the presence of many radio access advancements in similar district and the cross-over of the transmission inclusion of these organizations. The organization determination approach is the critical piece of the heterogeneous remote organizations. Well known network determination methods depend on exact qualities for network credits. Notwithstanding, due of variables including client versatility, remote sign impedance, and organization state incitation, the found organization properties are frequently uncertain. To take care of this issue, previously make use of multi-characteristic access choice methodology utilizing fuzzy organization credits. The system starts by ascertaining the beginning qualities for the stretch reluctant fuzzy hypothesis credits. From that point onward, the entropy approach is utilized to find the goal loads of organization property estimations, while the logical progressive system process is utilized to decide the abstract loads of organization trait values. The joined emotional and objective loads are gotten utilizing a method in view of the best mathematical distance to the pessimistic ideal arrangement. At last, competitor network scores were resolved involving dark social examination as per the intuitionistic fuzzy choice framework.

But it leads to research gap like Opportunistic and unmanageable best describe this system, Not a viable option for real-time use, Training the model requires a lot of computational resources, Require a lot of processing power and a sizable amount of memory, Flaws in the design of their own networks., etc. For better identifying the black hole attack in darknet, neural network [16] seems feasible. The services can be affected and many attacks are possible due

to breaching of data, non-availability of data, misinterpretation of data, denying the service, etc. [15]

3.PROPOSED ARCHITECTURE:

Proposed architecture is shown in fig-1For a more in-depth understanding of the data and its numerous facets, you might use a data analytics method known as exploratory data analysis. This helps us get a deeper appreciation for our data and spot previously hidden patterns.

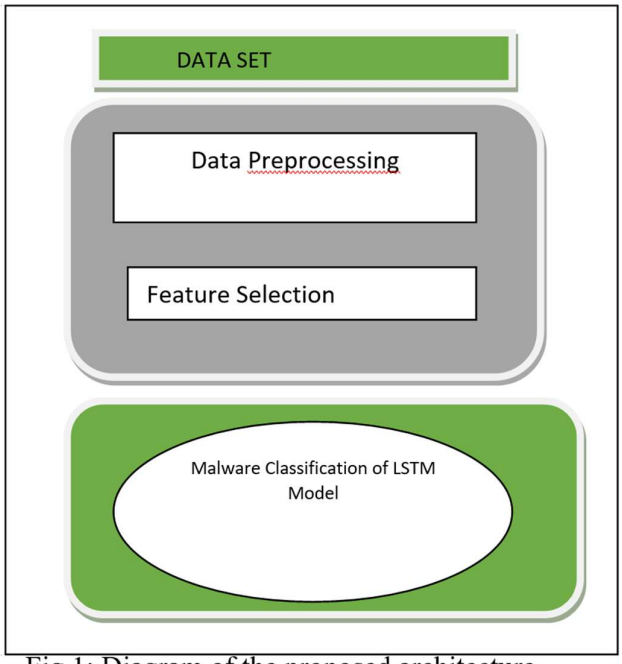

Fig 1: Diagram of the proposed architecture

One's primary objective should be data comprehension; when studying the data, understanding the data can mean many different things. While exploring the data, there are a few things to keep in mind, including checking for redundancy, missing values, or even null values on the data set. In order to choose which pertinent features/attributes to include and which irrelevant information to omit for predictive modelling, a series of techniques known as feature selection (FS) is used. It is an essential activity that helps machine learning classifiers lower error rates, decrease computation time, avoid overfitting, and increase classification accuracy.

The hidden state of the LSTM model is used to maintain information from inputs that have already passed. Since it stores sequential data in its hidden states, it functions well with sequential data. Since its inputs are limited to things observed in the past, the traditional LSTM model preserves historical information. Using the bi-directional LSTM, the model may be trained using inputs from the past as well as the future. This method may be used to save data for the future as well as the past. In essence, dissimilar to the ordinary LSTM model, which just gains setting from information sent in one far (ahead), the bidirectional LSTM model learns setting by moving information both forward and in reverse).

3.1Neural Network Long Short-Term memory (LSTM)

With its ability to remember past inputs, LSTM (Long Short-Term Memory) intermittent brain organizations (RNNs) are appropriate to handling successive information. The information

flow may be controlled on a case-by-case basis with the use of a memory cell, input door, yield entryway, and neglect entryway. The model is time-series backpropagation-prepared and can deal with input successions of variable lengths. Regular language handling, voice acknowledgment, and time series expectation are only a couple of the areas where LSTMs haveshown to be effective.

Using the model, it can be able to detect traffic and to categorize it which is shown in fig-2

Fig.2: Proposed model

3.2. Data Processing and visualization

Reducing memory utilisation and renaming columns are two examples of the data processing processes applied to the Darknet dataset. When working with huge datasets, minimising memory use is a frequent strategy for improving speed. In this situation, memory utilisation may be drastically decreased without any loss of information by using the "reduce_mem_usage" capability, which downcasts the information sorts of sections containing numeric qualities.

Traffic data set typewise is shown in below two figures 3(a) and 3(b)

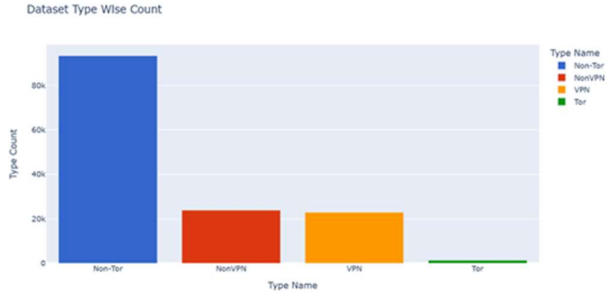

Fig.3(a): Classification of traffic dataset

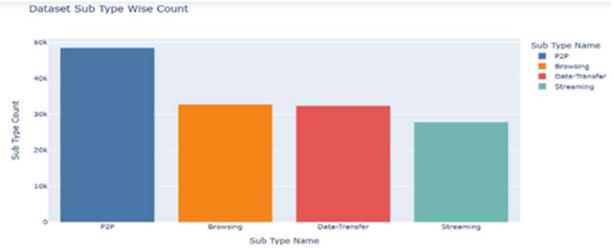

Fig.3(b): Classification of traffic dataset

3.3. Under-sampling and Feature Selection

After initialization, the code uses the imblearn library's RandomUnderSampler to do under-sampling. This is done when one class in the dataset has a disproportionately high number of samples compared to the other classes. To do this, the RandomUnderSampler picks a subset of samples at random from the majority class(es), therefore equating the sample sizes of the

minority classes.The code then uses the sklearn library's LassoCV and SelectFromModel to choose features to use. LassoCV is a linear regression model that uses L1 regularisation, which punishes excessively large model coefficients. As a consequence, the resulting solutions are sparse, with some coefficients equal to zero, a situation that may be used for feature selection. Any model having a "coef_" or "feature_importances_" property after fitting may be utilised with the SelectFromModel meta-transformer. It chooses features with coefficients greater than zero or feature importances greater than a certain threshold.

RandomUnderSampler:To ensure that each class has an equal number of instances, RandomUnderSampler will choose a certain number of samples at random from the larger class. In order to obtain the necessary ratio of minority class to majority class, a simple formula for RandomUnderSampler includes randomly picking samples from the majority class. Here is how the formula looks in text form:How Many Minority Participants = Target Proportion of Minority Participants to Total Participants * Total Participant Number

LassoCV:In order to identify features, the LassoCV linear regression model employs L1 regularisation. In order to promote sparsity in the coefficient values of the linear regression model, the L1 penalty is imposed to the coefficients, essentially setting certain coefficients to zero. LassoCV's formula calls for minimising the L1 penalty on the coefficients while minimising the number of squared blunders among anticipated and genuine upsides of the reliant variable.

The equation for the above is represented as

$$Min\left(\frac{1}{2 * nsamples}\right) * \left|\left|y - Xw\right|\right|^{\wedge}2 - 2 + \propto * \left|\left|\mathrm{w}\right|\right| - 1$$

Where the dependent variable's values are denoted by y and independent variable is denoted by X.The calculated coefficients are included in the vector w.

The L1 penalty's severity is determined by the regularisation parameter alpha, whereas n_ samples is the total number of data points.

SelectFromModel:

SelectFromModel is a feature selection technique for finding the most relevant aspects of a dataset by using a machine learning model[17]. SelectFromModel takes a machine learning model that has been fit to the dataset as input, then uses the model's feature significance scores to choose which features to use. This formula may be expressed as:Choose attributes that have significance ratings that are higher than or equal to a cutoffvalue.Machine learning models assign significance ratings to features based on how well they perform as predictors of the dependent variable.

3.4. LSTM Model Development

The given below sample code builds a Long Transient Memory (LSTM) brain network model for the point of completing a grouping position. The repetitive brain organization known as the long transient memory (LSTM) is helpful for foreseeing groupings in applications demanding time-series information or normal language processing. The model comprises of a 512-unit long momentary memory (LSTM) layer, 32-and 64-unit completely connected layers, and a 4-unit yield layer (because of the presence of 4 classes in the preparation information).

To ascertain class probabilities, the LSTM layer utilizes the relu enactment capability, though the result layer utilizes the softmax actuation capability. The Adam streamlining agent is utilized related to the straight out cross-entropy misfortune capability and the measurement of "precision" for estimating executionThe LSTM layer's input_shape specifies that the input data (X_train) be reshaped such that it contains 25 time steps and 1 feature. From that point forward, the model is prepared for 20 cycles on the X_train and y_train information with a bunch size of 256 and an approval split of 0.2.

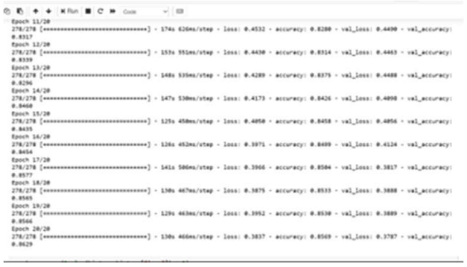

Fig -4(a) Sample code

```
In [113]: y_pred = model.predict(X_test)
          695/695 [==============================] - 44s 63ms/step

In [114]: y_pred = np.argmax(y_pred, axis=1)

In [115]: y_pred
Out[115]: array([0, 2, 3, ..., 1, 0, 0], dtype=int64)

In [116]: y_test
Out[116]: array([[1, 0, 0, 0],
                 [0, 0, 1, 0],
                 [0, 1, 0, 0],
                 ...,
                 [0, 1, 0, 0],
                 [1, 0, 0, 0],
                 [1, 0, 0, 0]])

In [117]: y_test = np.argmax(y_test,axis=1)

In [118]: y_test
Out[118]: array([0, 2, 1, ..., 1, 0, 0], dtype=int64)

In [119]: cm = confusion_matrix(y_test, y_pred)

          plt.figure(figsize=(7,5))
```

Fig-4(b)Sample Code

4.MODEL EVALUATION AND DEPLOYMENT:

A Long Short-Term Memory (LSTM) model is trained using the input training data (X_train and y_train). The model is then used to make predictions about the class marks of the test information (X_test), and these expectations are contrasted and the genuine names (y_test) from thetestingphase.Matplotlib is first used to display the loss and accuracy during training and validation to evaluate the model. This makes it simpler to evaluate the model's learning rate and spot signs of over- or underfitting while it is being trained. The code then generates the confusion matrix from the anticipated and actual labels, a table that compares the two to determine the classification model's efficacy. Seaborn is used to plot the confusion matrix, providing a visual representation of the model's efficacy. The diagonal members of the matrix

indicate the fraction of valid examples that were assigned to each class, while the non-diagonal elements stand in for the misclassified data. The overall execution of the model might be assessed utilizing various measures, like exactness, accuracy, review, and F1-score,all of which are calculated using the confusion matrix.

5. PERFORMANCE ANALYSIS:

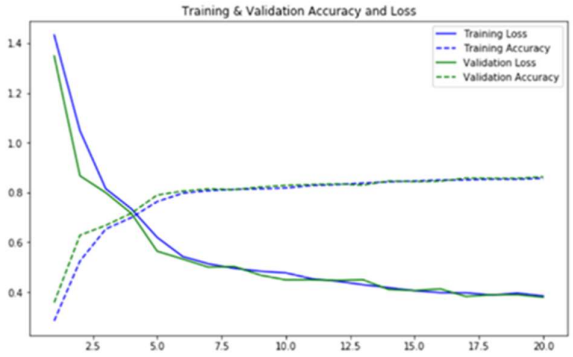

Fig-5-Performance Analysis

A cutting edge and broad dataset (i.e., CIC-Darknet) was utilized to assess the created DTDS-DL models. This dataset incorporates an enormous number of caught cyberattacks and unreported administrations given by the Darknet, and it was additionally characterized into four classifications (VPN, Peak, Non-VPN, Pinnacle). When contrasted with other cutting-edge models, the ongoing harvest of DTDS test systems benefits from our top discoveries too. Recommendations for the turn of events and future updates of this structure ought to focus on the robotized streamlining of the key technique pre-preparing boundaries to construct a much more effective, exact, and rapid arrangement process.

6.CONCLUSION AND FUTURE WORK:

It is challenging for businesses to proactively discover big security threats because attackers usually employ trusted, known technologies that are already deployed in network settings, and they swiftly adapt their evasion techniques. In light of the ever-evolving methods used by cybercriminals, the development of network traffic analysis tools has provided businesses with a practical means of countering these threats. Furthermore, successful organization perceivability has turned into an exceptionally extreme and complex methodology in light of the fact that to the broad utilization of distributed computing Gadget Administrators (DevOps) rehearses and the Web of Things (IoT). Network traffic examination innovation is especially significant as a result of its capacity to join its center gifts to permit malignant expectation location. In this examination, we make a model, build, and survey a completely independent Darknet traffic recognition framework and distribute our discoveries.

How this framework might be improved by involving a more complicated design with Siamese brain networks in an equal and conveyed setting or through blockchain is a significant area of study.

REFERENCES

[1] XiaoxueGuo,Mohd. HasbullahOmar,KhuzairiMohdZaini,GenLiang,Maoyuan Lin, ZirunGan, ‘‘Multiattribute Access Selection Algorithm for Heterogeneous Wireless Networks

Based on Fuzzy Network Attribute Values,'' IEEE Access., Vol. 10, pp. 74071-74081, 24 June 2022.
[2] A. Ahmed, L. Boulahia, and D. Gaiti, ''Enabling vertical handover decisions in heterogeneous wireless networks: A state-of-the-art and a classification,'' IEEE Commun. Surveys Tuts., vol. 16, no. 2, pp. 776–811, 2nd Quart., 2013.
[3] I. Modeas, A. Kaloxylos, L. Merakos, and D. Tsolkas, ''An adaptive and distributed network selection mechanism for 5G networks,'' Comput. Netw., vol. 189, Apr. 2021, Art. no. 107943.
[4] A. Keshavarz-Haddad, E. Aryafar, M. Wang, and M. Chiang, ''HetNets selection by clients: Convergence, efficiency, and practicality,'' IEEE/ACM Trans. Netw., vol. 25, no. 1, pp. 406–419, Feb. 2016.
[5] M. Kassar, B. Kervella, and G. Pujolle, ''An overview of vertical handover decision strategies in heterogeneous wireless networks,'' Comput. Commun., vol. 31, no. 10, pp. 2607–2620, Jun. 2008.
[6] E. Obayiuwana and O. E. Falowo, ''Network selection in heterogeneous wireless networks using multi-criteria decision-making algorithms: A review,'' Wireless Netw., vol. 23, pp. 2617–2649, Nov. 2017.
[7] R. Trestian, O. Ormond, and G.-M. Muntean, ''Performance evaluation of MADM-based methods for network selection in a multimedia wireless environment'' ,WirelessNetw., vol. 21, no. 5, pp. 1745–1763, Jul. 2015.
[8] Y. Zhong, H. Wang, and H. Lv, ''A cognitive wireless networks access selection algorithm based on madm,'' Ad Hoc Netw., vol. 109, Dec. 2020, Art. no. 102286.
[9] D. Jiang, L. Huo, Z. Lv, H. Song, and W. Qin, ''A joint multi-criteria utilitybased network selection approach for vehicle-to-infrastructure networking,'' IEEE Trans. Intell. Transp. Syst., vol. 19, no. 10, pp. 3305–3319, Jan. 2018.
[10] X. Wu and Q. Du, ''Utility-function-based radio-access-technology selection for heterogeneous wireless networks'',Comput. SElectr. Eng., vol. 52, pp. 171–182, May 2016.
[11) Lira-YOLO: a lightweight model for ship detection in radar images ZHOU Long, WEI Suyuan, CUI Zhongma1, FANG Jiaqi1, YANG Xiaoting, and DING Wei, Journal of Systems Engineering and Electronics Vol. 31, No. 5, October 2020, pp.950– 956
[12] SeunghyunYoon Jin-Hee Cho , Dong Seong Kim, Terrence J. Moore, Frederica Free-Nelson, and Hyuk Lim , Attack Graph-Based Moving Target Defense in Software-Defined Networks , IEEE TRANSACTIONS ON NETWORK AND SERVICE MANAGEMENT, VOL. 17, NO. 3, SEPTEMBER 2020
[13]Tetsuya Shigeyasu and AyakaSonoda,"Detection and mitigation of collusive interest flooding attack on content centric networking", International Journal of Grid and Utility Computing,Vol. 11, No. 1,2020, pp 21–29,doi.org/10.1504/ijguc.2020.103966
[14]Ru Huang , Lei Ma , GuangtaoZhai , Jianhua He , Xiaoli Chu , Senior Member, and Huaicheng Yan1 , " Resilient Routing Mechanism for Wireless Sensor Networks With Deep Learning Link Reliability Prediction", IEEE Access, Vol-8,Page(s): 64857 – 64872,2020,DOI: 10.1109/ACCESS.2020.2984593
[15] M. S. Bennet Praba and J S FemildaJosephin, Review on various authentication schemes and attacks on connected vehicles, : 2020 IOP Conf. Ser.: Mater. Sci. Eng. 993 012102

[16]Vigil, A., Bharathi, S.”Diagnosis of pulpitis from dental panoramic radiograph using histogram of gradients with discrete wavelet transform and multilevel neural network techniques” Traitement du Signal,2021, 38(5), pp. 1549–1555
[17]Kumar, R., Sodwadia, D., Antony Vigil, M.S., Ghosh, S., “Increasing efficiency and prediction of heart disease using machine learning algorithms”, International Journal of Advanced Science and Technology, 2020, 29(3), pp. 5985–5991
[18]Chelliah, B.J., Antony Vigil, M.S., Bennet Praba, M.S.”Node clone detection using a stable overlay network”,International Journal of Electrical and Computer Engineering, 2020, 10(1), pp. 316–322