International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Decision Tree Based Data Pruning with the Estimation of Oversampling Attributes for the Secure Communication in IOT

Dr. Nachaat Mohamed[1], Mr. Aishwary Awasthi[2], Dr Nandini Kulkarni[3], Dr. Sridhar Thota[4] Mandeep Singh[5], Sumedh Vithalrao Dhole [6]

Submitted: 22/08/2022 **Accepted:** 26/11/2022

Abstract: Internet of Things (IoT) exhibits a significant role to evaluate the error or supply shortage. The IoT demand for the security and authentication of the devices is considered as the most priority for software developments. As the IoT communication comprises of an interconnected environment for the both digital and physical scenarios. The IoT environment exhibits anything and anywhere services to the communication medium. In those scenarios, security is considered as the major concern to protect the data resources from unauthorized resources for the appropriate security and privacy. This paper proposed a decision tree-based pruning scheme for the IoT attributes. The proposed decision tree based pruning for the security attributes are defined as the decision tree pruning (DTP). The proposed DTP model comprises of the minority oversampling model for the estimation of the attack features. With the developed DTP model, the attack datasets were pre-processed and evaluated for the different attack environments in to consideration. The DTP processed data were applied over the conventional machine learning-based model for the computation attacks in the network. The simulation results expressed that proposed DTP model achieves the accuracy value of 98% which is ~3% higher than the conventional classifier techniques.

Keywords: *Security, Attacks, Internet of Things (IoT), Pruning, Decision Tree, Oversampling Model*

1. Intoduction

Security is the central issue in a computing system that is connected to the Internet of Things (IoT) which will be in the form of active or passive attacks. The security and privacy issue in IoT and the need for security challenges due to security and privacy issues are also present. Briefly explained the detection of attacks by using machine learning methods [1]. IoT is the network of physical data such as smartphones, vehicles, home appliances, building, and other objects that are embedded with sensors, electronic software, and connectivity in a network that improves the devices to interchange and collect the data. The IoT has the capacity to make the object sense, recognize and control it remotely by the infrastructure of an existing network. The properties of IoT create the opportunity to incorporate in the physical world on a computer-based system which effectively enhances the accuracy, efficiency, and economic benefit [2]. When the IoT is combined with actuators and sensors, it becomes a cyber-physical model that also uses the technologies such as intelligent transportation systems, smart homes, smart grids, and smart cities. Every object is identified uniquely by using an embedding computing system that enables it to interoperate among the infrastructures of the internet. IoT is the innovative paradigm that consists of billions of things that intelligently communicate [3].

Recently, the IoT became an emerging domain in various sectors of applications. With the emerging technologies, a vast number of people are connected to the internet by utilising mobile phones, computers, and laptops. IoT includes various directions for the research field, which leads to the design and development of a real-world application for society [4]. IoT is also defined as the next generation of the internet or the future of the internet, where billions of objects or things will be interacting with each other. The IoT is the collection of things that consists of objects that include different identities over the internet [5].

Security is an essential requirement in the environment of IoT because the IoT is highly vulnerable to attacks for various reasons. Initially, the IoT devices are increasing in number, and the devices will be located in an unmanaged and managed environment [6]. So, the attackers will be able to utilise the unmanaged devices at a remote place to generate the attacks. However, the connected devices in the physical environment of users will cause damage if it is compromised. Then, the IoT includes heterogeneous devices and technology to develop the issue of interoperability which shows the way for attackers [7]. Further, the impacts of attack on the environment of IoT go beyond certain devices, such as the compromised devices are utilised in botnets to generate massive attacks. So, the attacks need to be neutralized and detected effectively in order to reduce the impacts. The existing solutions are certain to the identification mechanism or mitigation process in IoT [8]. But the autonomous security provider methods are limited in number and are unsuitable in the environment of IoT for new attacks that are rising due to the increasing number of IoT devices. So, the increasing number of IoT devices need a protection system that automatically identifies known and new attacks and defends the detected attack at a higher rate. This research aimed to increase

[1] *Rabdan Academy, UAE, Abu Dhabi, Homeland Security Universiti Sains Malaysia, Abu Dhabi, eng.cne1@gmail.com*
[2] *Research Scholar, Department of Mechanical, Sanskriti University, Mathura, Uttar Pradesh, India, aishwary@sanskriti.edu.in*
[3] *Symbiosis School of Planning Architecture and Design, Symbiosis International Deemed University., Symbiosis School of Planning Architecture and Design, Nagpur, deputydirector@sspad.edu.in*
[4] *Professor, Department of ECE, Alliance College of Engineering and Design, Alliance University, Bangalore, India sridhar.t@alliance.edu.in*
[5] *Assistant Professor, Department of Physical Education, University of Jammu, Jammu, mandeep.singh@jammuuniversity.ac.in*
[6] *Assistant Professor, Department of ECE, Bharati Vidyapeeth (Deemed to be University), College of Engineering,Pune-satara road, Dhankawadi, Pune, Maharashtra, India.*, svdhole@bvucoep.edu.in

the security in the IoT communication with an effective security model for the attack classification with estimation of attributes.

2. Related Works

In [9] presented identification of Denial of Service (DoS) attack using Message Queuing Telemetry Transport (MQTT) protocol. Additionally, the size/length feature sets of the MQTT control packet field were examined for two different datasets, such as [10] and [11]. Here, three different machine learning methods were used to analyze the capabilities of feature sets such as typical estimator-based NB, C4.5 based DT, and ANN. In the basic access of MQTT broker, a huge impact was created, as shown in the modelling of MQTT DoS attack. An inappropriate subscription flooding attack also impacted IoT devices. But illegal authentication attacks made less impact on a single attack source machine which was mainly based on a considerable volume of attack packets. If features related to control packet field size length were chosen, then the DoS detection model showed higher detection capacities of MQTT features. Therefore, the distribution features of field length and packet size were utilized during DoS attack detection. However, utilization of higher memory occurred based on the usage of abnormal CONNECT requests, which were used while accomplishing the memory exhaustion attacks. In [12] developed an intelligent architecture that integrated Machine Learning (ML) and Complex Event Processing (CEP). It has the capacity of simply handling dynamic patterns for identifying security attacks in the IoT system. The dynamic pattern was comprised of some event properties, which were based on automatically generated values obtained from the support vector regression and linear regression predictor. The major contributions of this work were separated into two phases. In the first phase, the ML technique combined with service-oriented architecture is used. This simplifies real-time identification and protection of attacks in IoT systems. The security attack's graphical definition was supported by using an extended MEdit4CEP. These security attacks were identified and prevented in the developed architecture. Seven types of attacks were considered in this IoT system, Discwave attack and Subfuzzing attack. The developed intelligent architecture failed to obtain an automatic prediction of features. Also, lesser training of data led to lesser accuracy.In [13] developed a security design and combined procedure for detecting cyber-attacks which was used to accomplish security in IoT. This work used the Bot-IoT dataset that was a widely obtainable operation for identifying Bot-IoT. The Bot-IoT dataset has various attacks. At first, the dataset of BoT-IoT identification was used, and its traffics were identified in the IoT system. Then, 44 effective features were selected from the dataset for the machine learning algorithm. The main objective of this work was to select an appropriate ML method as well as to recognize anomaly intrusion in IoT network traffic. Here, five different ML algorithms are considered - random forest, Random Tree, NB, C4.5 DT, and Bayes Net. Subsequently, a mathematical model, namely the bijective soft set technique, was used to select an appropriate ML method from various ML methods. Finally, a hybrid ML algorithm selection was developed to achieve intrusion and anomaly detection in the IoT network traffic. Chosen ML algorithm was used to achieve the effective classification of the intrusion and anomaly IoT traffic. But this security framework model failed to identify the IoT anomalies while detecting the traffic.

In [14] developed network-based IDS to monitor malicious traffic in the IoT network. In this work, an IoT network was constructed by using edge and fog devices. This IDS model followed both distributed and decentralized (semidistributed) models and developed ML-based IDS, using possible fog-edge collaborative analytics. The unknown attacks were detected using anomaly-oriented detection, and these attacks were prevented from obtaining normal operations. Leveraging broad computational tasks, from single powerful units to multiple less powerful units, was 42 enabled using the IDS model. This resulted in generating an impressive IDS model for IoT networks that comprised an enormous amount of edge and fog devices. For the instantaneous selection of features, parallel models were operated on the edge side in the semi-distributed method. Then, output features were processed using a single Multi-Layer Perceptron (MLP), which was a course on the cloud. The corresponding models independently accomplished both MLP classification and selection of features in the distribution method. The final decision-making was achieved by integrating parallel outputs accessed in the coordinating edge or fog. Higher detection accuracy was achieved in the semi-distributed method with higher building due to collective classification features. In the distributed method, lower accuracy was achieved by higher processing because of individual classification and processing. This work used Aegean Wireless Intrusion Detection (AWID) dataset for the IDS model in the IoT, since AWID cyberattack dataset has the data of all the previous records. However, the distributed approach can design a computational model that will be 2.5 intervals quicker over a semidistributed one. However, this network-based IDS detection achieved only partial elimination of temporal features rather than removing them completely.

3. Decision Tree Pruning with the Over Sampling in IoT Security

The DTP model performs the with decision tree process to minimize the complexity associated with the security attributes in the IoT. The proposed DTP model comprises of the subtree leaves in the node for the reduction of the data size. The data size in the proposed DTP is minimized with the pruning process and accuracy is increased for the attack classification. With the proposed DTP mode the feature values are estimated to eliminate the imbalance classification. The dataset imbalance subjected to the challenges in the machine learning process with the minority classes leads to reduced performance. The DTP uses the oversampling approach for the minority class for every sample class through line segmentation with the nearest neighbor estimation.

The attacks in the IoT environment with the proposed DTP process is evaluated with the binary class estimation as normal or abnormal class. The multi-class attacks are estimated with the NSL-KDD dataset with consideration of the 4 attack scenario such as DoS, Probe, R2L, and U2R. The DTP advantage comprises of the linear computational complexity process for the significant training process. With the proposed DTP model the error values are computed for the each nodes and error cost in the nodes are computed as in equation (1)

$$R_c(t) = r_m(t) \times r_e(t) \quad (1)$$

Where, the DTP error cost are defined as $Rc(t)$ and the other features are presented as follows:

$$r_m = \frac{Number\ of\ Misclassification\ in\ the\ nodes}{Number\ of\ example\ nodes}$$

$$r_e = \frac{Numebr\ of\ example\ nodes}{Number\ of\ total\ samples}$$

With the proposed DTP model each node compute the values of the node *t* with the pruned value with computation of the error cost and the rooted values are defined as in equation (2)

$$R_c(T) = \sum_{i=number\ of\ nodes} R_c(i) \quad (2)$$

The process involved in the DTP model for the attack classification algorithm are presented as follows:

Algorithm 1: Pseudo Code for DTP

Initially, decision tree is constructed
for
    each node

if node is parent, then
    The risk rate of the node as *(RI)* and risk rate of parent as *(Rp)*
If ($Rp > Rl$)
If parent node is converted as the leaf nodes
Estimate the error cost using the equation (4)
end if
end if

end for
Return

With the proposed DTP model the overall process presented in figure 1.

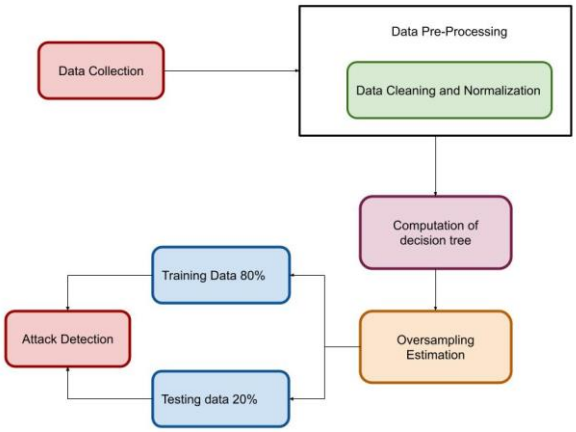

Figure 1: Process Flow in DTP

3.1 Pre-Processing and Splitting Stage

Once the data acquisition process from specified datasets is done, preprocessing is established by means of the data normalization and cleaning process. Data cleaning is one of the methodologies of adjusting and identifying the incorrect or tainted reports commencing from the data set. Additionally, it tracks down the mistaken, erroneous, deficient, or unessential pieces of the information and afterwards supplants, alter or erases the coarse information as in equation (3)

$$n_i = \frac{a_i - min_{ai}}{max_{ai} - min_{ai}} \quad (3)$$

where the normalizing value of the $i\ th$ variable is stated as a_i , minimum & maximum valour registered for i^{th}variable is designated as $min_{ai}$and max_{ai}. Once the above-stated two stages for precision improvement of RF-DTP are completed, selection of best features from the dataset for the classification of attacks is carried out using Recursive Feature Elimination (RFE) technique.

3.2 Classification of IoT Attacks with DTP model

Once the data splitting process is completed, DTP is introduced to categorize the traffics types in both the specified datasets. RF is considered as an operative classifier for large datasets. The reason is that it exploits superior data standards for the classification of various attacks and produces numerous decision trees (DTs), and syndicates the trees organized for progressing towards an effectual classifier. By using information gain and entropy values the decision tree t algorithm selects the most significant features for branching.

The entropy value and information gain are calculated using the equation (4) – (6) as follows:

$$Gain(t,x) = E(t) - E(t,x) \quad (4)$$

$$E(t) = \sum_{i=1}^{c} -P_i log\ _2 P_i \quad (5)$$

$$E(t) = \sum_{c \in n_i} p(c) E(c) \quad (6)$$

where the class is stated as c, entropy features of x are designated as $E(t, x)$, and class entropy are declared as $E(t)$. Here, respective tree t is deliberated equally as a distinctive classifier which is exploited to accomplish an improved decision-making process. The error speed of the Random Forest classifier rests on two factors; the relationship among the trees ought to be small, and the strength of the tree would be large for diminishing the error rate. The DTP technique, thus, successfully employed to balance the normalized data vectors ni, which supports the Random Forest classifier to accomplish improved outcomes in the classification of traffic. The effective oversampling technique is the DTP, which instead of data replacement.

4. Performance Analysis

Here, the dataset NSL-KDD is commenced for authenticating the presentation of the suggested DTP. Here, 125973 and 22544 records are used for the process of training and testing. Here, the presentation of the binary classification is presented in the respective Table 1 and Table 2. From the tables, the performance of the projected RFDTP method is evaluated through f-measure precision and recall. From Table 3, the projected DTP (Random Forest and Synthetic Minority Oversampling Technique) accomplished 98.61%, 98.41%, and 98.56% of precision, recall, and f-measure significance, respectively.

Table 1: Comparison of Classifier

Classifier	F-measure (%)	Precision (%)	Recall (%)
SVM	88.67	89.56	84.36
RF	89.77	89.83	90.13
Decision tree	90.34	90.69	90.94
DTP	97.89	98.67	98.93

The figure 2 provides the performance comparison of the proposed DTP with the existing classifier with the proposed DTP model is presented.

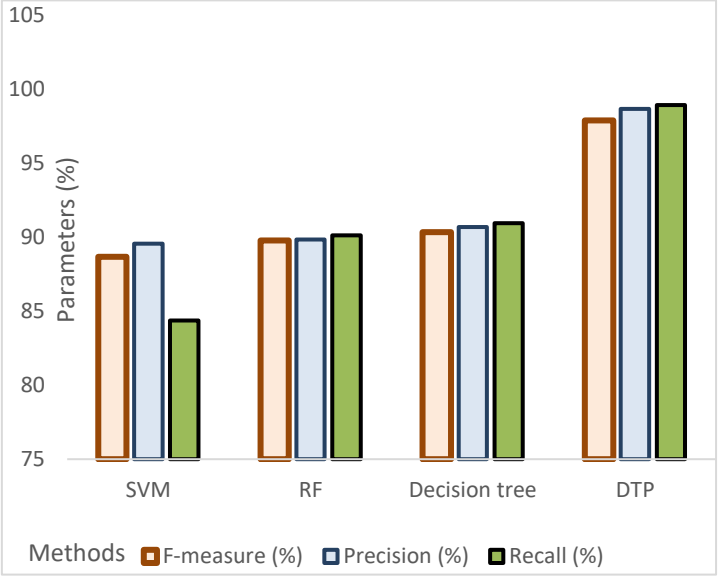

Figure 2: Comparison of Parameters

The performance of DTP in Table 2 is analyzed in terms of AUC, accuracy, FAR compared through the supporting methods as mentioned in the Table. By examining Table 2, proposed -DTP standardly accomplished accuracy as 98.31%, FAR as 0.21%, and AUC as 99.87%, which is much improved while associated with conventional methods. Thus, from the analysis, the proposed DTP model exhibited improved results when compared with existing methods.

Table 2: Comparison of NSL-KDD dataset for accuracy

Classifier	AUC (%)	FAR (%)	Accuracy (%)
SVM	93.45	0.32	89.74
RF	94.46	0.34	92.35
Decision tree	95.78	0.42	94.67
DTP	98.53	0.19	99.35

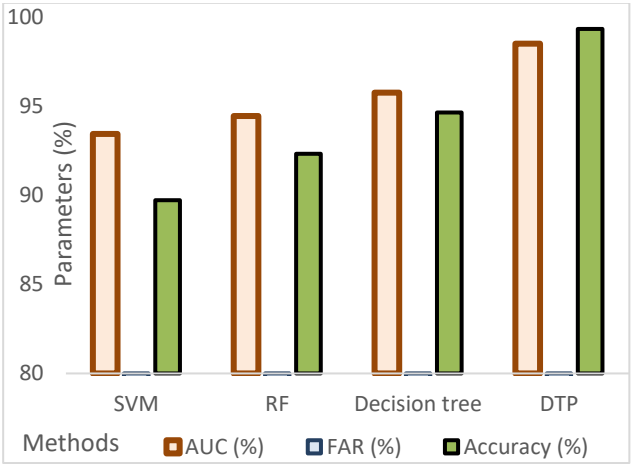

Figure 3: Comparison of performance for the NSL-KDD

In figure 3 the comparative analysis of the proposed DTP model for the different attributes of the IoT data is presented. The presentation analysis of 4-class for the NSL-KDD dataset is furnished in Table 3, respectively. From Table 3, DTP results are analysed using precision, f-measure, and recall. And RF-DTP averagely accomplished precision as 99.01%, f-measure as 98.94%, and recall as 98.33% that is much improved while comparing with remaining methods as mentioned in [4]. The outcome of precision, fmeasure and recall, on NSL-KDD is presented in Figure 6.

Table 3: Overall Performance Analysis for different attacks

Classifier	Class	F-measure (%)	Recall (%)	Precision (%)
SVM	DoS	86.56	84.88	93.44
	Probe	84.33	89.48	88.62
	R2L	66.11	66.09	85.31
	U2R	73.27	73.20	87.20
Decision Tree	DoS	85.41	84.71	88.77
	Probe	95.51	94.24	95.79
	R2L	96.77	96.72	96.83
	U2R	96.57	97.63	96.46
RF	DoS	96.81	96.71	95.92
	Probe	97.82	97.71	97.86
	R2L	87.13	88.21	95.64
	U2R	95.79	94.55	96.16
DTP	DoS	99.67	98.74	99.75
	Probe	98.97	98.64	98.59
	R2L	99.39	99.29	99.54
	U2R	99.62	99.74	99.98

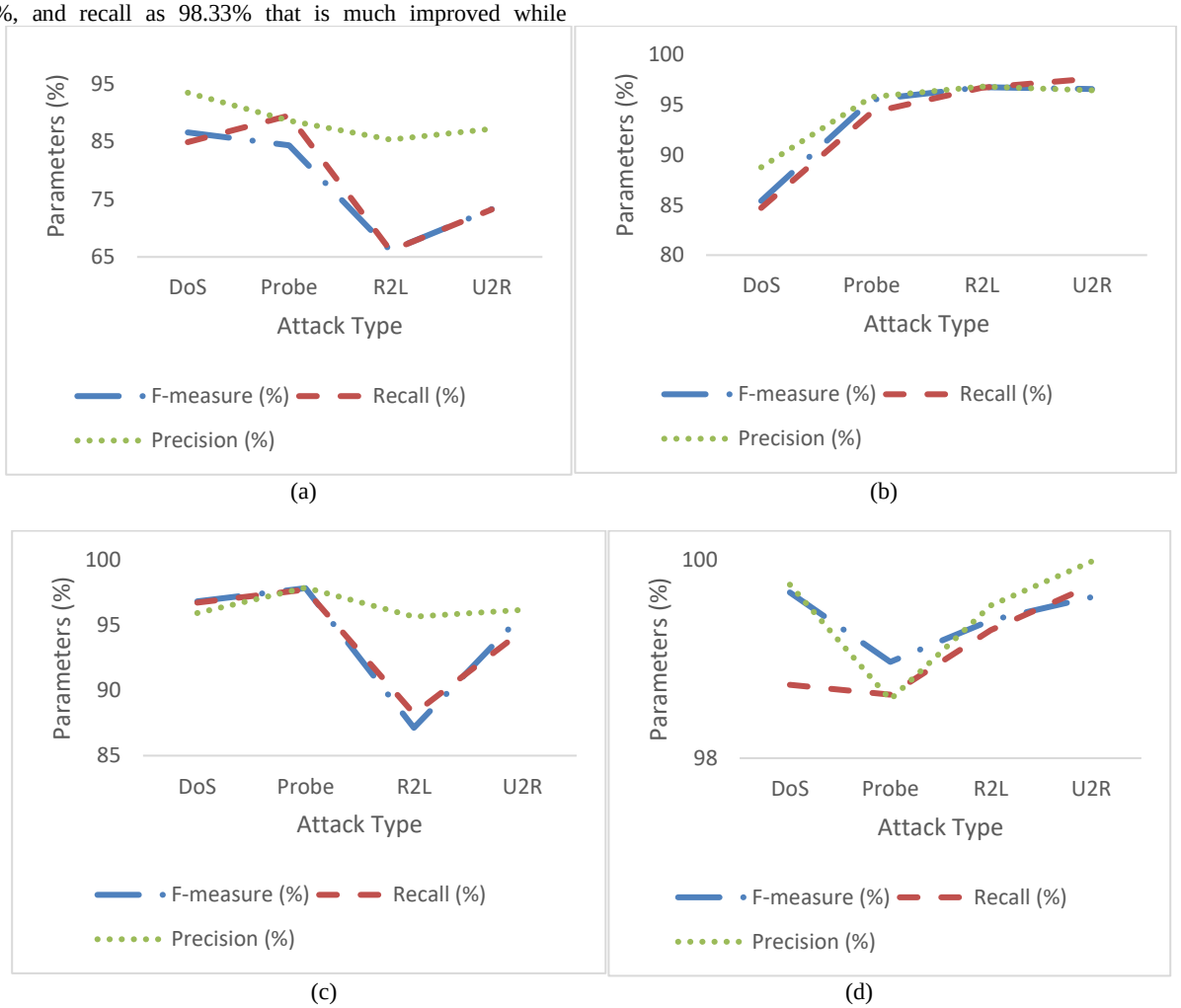

Figure 4: Comparison of performance for different attacks (a) SVM (b) Decision Tree (c) Random Forest (d) DTP

Here, the RF classifier is exploited for emerging well-organized and operative IDS. To enhance the detection rate of all the attacks in the imbalanced dataset DTP is introduced, which is also chosen for every significant structure of the minority class by means of the attack method. Outcomes commencing from the research, displays that the DTP method increased the detection rate in a significant way. Random Forest with the DTP technique is projected in order to identify the attacks. Thus, the assessment is conceded on N-BaIoT and NSL-KDD datasets for the detection of attacks. The assessment outcome displays that the proposed DTP successfully distinguishes the normal and IoT attacks traffic using f-measure, recall, precision, AUC, accuracy, and FAR. The projected DTP

presented the least of 0.14% and an extreme of 14.25% enhancement in accuracy on binary class. Furthermore, DTP standardly presented minimum and maximum of 0.04% and 7.35% development in accuracy on four class. Conversely, the suggested DTP (Random Forest and Synthetic Minority Oversampling Technique) exhibited the least possible value & maximum value of 0.01% and 0.04% enhancement on N-BaIoT when associated with conventional classifiers, for example, deep model, decision tree, SVM, shallow model, and RF. The attained outcome revealed that the proposed DTP delivers improved distinguishability of IoT attacks.

5. Conclusion

As the IoT is the emerging technology which offers the effective service delivery to the users. However, the implementation of the IoT subjected to the challenge associated with the security concern. The security is major challenge for the wireless IoT system involved in the authenticated data transmission between the nodes. This paper presented a DTP model to provide secure data transmission between the nodes. The proposed DTP model uses the decision tree with the oversampling reduction technique in IoT application. The proposed model performance is evaluated for the consideration of the different attacks in to consideration. The comparative analysis expressed that proposed DTP model exhibits the higher accuracy rate of 99.98% which is significantly higher than the other attack classifiers. The analysis confirmed that proposed DTP model exhibits higher security compared with the existing classifier models.

References

[1] Shaukat, K., Alam, T. M., Hameed, I. A., Khan, W. A., Abbas, N., & Luo, S. (2021, September). A review on security challenges in internet of things (IoT). In 2021 26th International Conference on Automation and Computing (ICAC) (pp. 1-6). IEEE.

[2] Yu, Z., Song, L., Jiang, L., & Sharafi, O. K. (2021). Systematic literature review on the security challenges of blockchain in IoT-based smart cities. Kybernetes.

[3] Thoutam, V. (2021). Unique Security Challenges Of Iot Devices And Spectrum Of Security Considerations. Journal of Artificial Intelligence, Machine Learning and Neural Network (JAIMLNN) ISSN: 2799-1172, 1(02), 1-7.

[4] Touqeer, H., Zaman, S., Amin, R., Hussain, M., Al-Turjman, F., & Bilal, M. (2021). Smart home security: challenges, issues and solutions at different IoT layers. The Journal of Supercomputing, 77(12), 14053-14089.

[5] Marshal, R., Gobinath, K., & Rao, V. V. (2021, April). Proactive Measures to Mitigate Cyber Security Challenges in IoT based Smart Healthcare Networks. In 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS) (pp. 1-4). IEEE.

[6] Raghuvanshi, A., Singh, U. K., Shuaib, M., & Alam, S. (2021). An investigation of various applications and related security challenges of Internet of things. Materials Today: Proceedings.

[7] Malhotra, P., Singh, Y., Anand, P., Bangotra, D. K., Singh, P. K., & Hong, W. C. (2021). Internet of things: Evolution, concerns and security challenges. Sensors, 21(5), 1809.

[8] Karale, A. (2021). The challenges of IoT addressing security, ethics, privacy, and laws. Internet of Things, 15, 100420.

[9] Mishra, N., & Pandya, S. (2021). Internet of things applications, security challenges, attacks, intrusion detection, and future visions: A systematic review. IEEE Access, 9, 59353-59377.

[10] Mohanty, J., Mishra, S., Patra, S., Pati, B., & Panigrahi, C. R. (2021). IoT security, challenges, and solutions: a review. Progress in Advanced Computing and Intelligent Engineering, 493-504.

[11] Balogh, S., Gallo, O., Ploszek, R., Špaček, P., & Zajac, P. (2021). IoT Security Challenges: Cloud and Blockchain, Postquantum Cryptography, and Evolutionary Techniques. Electronics, 10(21), 2647.

[12] Azrour, M., Mabrouki, J., Guezzaz, A., & Kanwal, A. (2021). Internet of things security: challenges and key issues. Security and Communication Networks, 2021.

[13] Ali, R. F., Muneer, A., Dominic, P. D. D., Taib, S. M., & Ghaleb, E. A. (2021, August). Internet of Things (IoT) Security Challenges and Solutions: A Systematic Literature Review. In International Conference on Advances in Cyber Security (pp. 128-154). Springer, Singapore.

[14] Gupta, H., & Sharma, S. (2021, June). Security Challenges in Adopting Internet of Things for Smart Network. In 2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT) (pp. 761-765). IEEE.