International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 11 | Nov - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Machine Learning Algorithm as a Firewall Decision and Reinforcement in Market Segmentation and Big Data

Aditya Singh[1], Siddharth Nag[2], Sakshi Pateriya[3], Ashok Masih*[4], Aakrati Nigam[5]
Shri Ram Swaroop Memorial University

Abstract
In the modern digital landscape, vast amounts of data are generated daily, challenging traditional analytical approaches in both data security and market segmentation. This paper presents a machine learning (ML) algorithm designed to serve a dual purpose: as a decision-support firewall system and as a reinforcement tool for market segmentation within big data environments. By integrating supervised and unsupervised learning models, the proposed algorithm effectively detects anomalies and potential security threats while also identifying distinct customer segments with high precision. The firewall decision-making component utilizes predictive models to detect malicious activity in real-time, enhancing cybersecurity by proactively learning from previous threat patterns. Concurrently, the reinforcement learning component analyses customer behaviour and preferences, dynamically adapting segmentation strategies to maximize marketing effectiveness. The dual implementation of ML in these domains demonstrates significant potential in improving both data security and personalized marketing outcomes. Experimental results indicate enhanced firewall accuracy and a refined segmentation process, suggesting that the proposed ML model provides a comprehensive solution for the challenges posed by big data in cybersecurity and market segmentation. s.

Keywords
Machine Learning; Firewall Decision; Reinforcement Learning; Market Segmentation; Big Data

INTRODUCTION

“Bigdata is the result of technological advancements in all facets of human existence. Massive volumes of data are generated by mechanical, IoT, and human devices.One Internet-based data transmission is susceptible to a variety of possible invasions and cyberattacks. Hackers could alter the accuracy and consistency of network data over its whole life cycle and transfer the data to unauthorised parties once the network infrastructure has been compromised. As a result, numerous safety techniques, such Internet firewalls, have been employed throughout the various Défense stages to address security issues.2. The internet is always expanding at an exponential rate. Because of this significant expansion, there is now a risk from cyberattacks. Cyber dangers have significantly increased in businesses across all industries throughout the last ten years. Organisations are at risk from cyberthreats like ransomware, phishing, data leaks, hacking, and insider threats. This indicates that procedures must be put in place to safeguard the usability and integrity of data.3. In high-performance computing (HPC) settings, network security is essential, especially when supercomputing services are being offered. Configuring and using security devices is a crucial part of the HPC system. In order to detect cyberattacks, collect logs from target nodes, and gather events from network security devices for anomaly detection, many of these security devices make use of independent log-collection and analysis servers.4. Everything in our life is digitally recorded, and we live in the age of data,

when everything is connected to a data source.5. The Internet of Things (IoT) takes into account how various objects are connected to one another, such as mechanical systems, intelligent sensors, autonomous cars, mechanisms and terminals, and industrial systems. Massive volumes of data have been produced as a result of the quick development of technology, which has created opportunities as well as difficulties for various businesses. Big data analytics is becoming a vital component of strategic decision-making, giving businesses important insights into the preferences, behaviour, and new trends of their customers. But this enormous increase in data also poses a serious cybersecurity threat, thus safeguarding data assets is of utmost importance."

"The integration of machine learning (ML) in addressing these dual challenges – securing data and utilizing it effectively for market segmentation – has become a promising approach. Machine learning techniques can automate complex processes, detect anomalies in real-time, and adapt dynamically to new patterns, providing more efficient solutions for data security and targeted marketing."

"This paper proposes a machine learning algorithm that functions as both a firewall decision-making mechanism and a reinforcement tool in market segmentation. The firewall aspect of the algorithm utilizes predictive modelling to identify and block malicious activities, thereby safeguarding sensitive information. On the other hand, the market segmentation aspect leverages unsupervised learning methods to analyse consumer behaviour patterns, allowing companies to effectively target specific segments and optimize their marketing campaigns. By integrating supervised learning for cybersecurity and reinforcement learning for market segmentation, the proposed solution aims to bring together two critical domains that are often handled separately. The concept of using machine learning for firewall decision-making draws upon its ability to learn from historical attack data and continuously improve threat detection capabilities, minimizing false positives and enabling a more proactive approach to cybersecurity. Simultaneously, the reinforcement aspect allows market segmentation models to adapt to changing consumer preferences, enhancing the ability to engage with customers on a personal level."

This introduction outlines the necessity of combining machine learning for firewall decisions and market segmentation, focusing on how big data analytics can be leveraged for dual purposes: securing information and deriving actionable insights for business growth. The proposed ML algorithm aims to address the intricacies of managing big data in these contexts, offering an integrated solution for both security and strategic marketing challenges.

"The following are the Research Primary Contribution"

Dual-Function Machine Learning Algorithm Development of a novel machine learning algorithm that serves a dual purpose: acting as a decision-support firewall for cybersecurity and providing reinforcement in market segmentation within big data environments.

Firewall Decision-Making Mechanism Implementation of a supervised learning approach to create a proactive firewall decision-making model capable of detecting malicious activities in real-time, thereby enhancing data security through predictive modelling and continuous learning from historical threats.

Dynamic Market Segmentation Utilization of unsupervised and reinforcement learning techniques to improve the efficiency of market segmentation, enabling dynamic adaptation to consumer behaviour patterns and preferences, which results in more targeted and effective marketing strategies.

Integrated Approach to Data Security and Market Analysis Presentation of a unified framework that simultaneously addresses data security challenges and big data market analysis, demonstrating how machine learning can be applied to secure data assets while also deriving meaningful insights for business growth.

"Enhanced System Performance Demonstration of significant improvements in both firewall accuracy and customer segmentation quality, highlighting the effectiveness of combining machine learning models for both anomaly detection in cybersecurity and pattern recognition in consumer behavior analysis."
Real-Time Adaptability Introduction of real-time adaptability features that enable the machine learning model to learn continuously from new data, ensuring its relevance and robustness in both detecting evolving cyber threats and adapting to changing market dynamics.

Scalable Big Data Solution A focus on scalability and efficiency, ensuring that the proposed machine learning algorithm can handle large datasets typical in big data environments, making it applicable for enterprises seeking both secure and insightful use of their data.

"Section I, the introduction, gives a summary of the research topic and establishes its context and relevance; Section II, related work, examines previous research and literature in the field to identify any gaps or connections with the current study; Section III, the problem statement, clearly states the issue or gap that the research is trying to fill; Section IV, methodology, describes the methods and strategies used to carry out the study; Section V presents the findings and has a discussion."

RELATED WORKS

"*Khoshaba Farah et al.* [1] With the aid of artificial intelligence (AI) and machine learning (ML), a vast array of diverse, heterogeneous, and even divergent data sources have been incorporated into computer science research concepts, yielding exceptional accuracy and data quality outcomes. Nevertheless, it is computationally and logistically expensive to apply machine learning techniques to highly large and dynamic datasets. Because certain people constantly generate large amounts of data, powerful analytics tools are more important than ever. A spark Regression, grouping, dimension reduction, and rule extraction are just a few of the many ML tasks that machine learning can perform. Despite this, it contains a large number of high-quality datasets with consistent organisation. The enormous density of large-volume processing was the target of the MapReduce programming model specification. By dividing the task into multiple clearly defined sub-sized components, it achieves this. Different data mining frameworks can be connected to a distributed device. The distributed system's infrastructure is provided by Hadoop. Mahout will scale to accommodate large datasets, according to a number of sources. In addition to discussing and illustrating the application of algorithms and their noteworthy influence on Big Data, this article compares the ways in which Apache Spark and Apache Mahout implement them."

"*Qasem Abu Al-Haija* [2] Using specific decisions to improve cyber-defense and filter out malicious packets, a firewall system ensures traffic control for both incoming and outgoing packets moving over communication networks. The filtration procedure compares the traffic packets to pre-established rules in order to prevent cyber threats from entering the network. The firewall system will either "accept," "deny," or "drop/reset" the incoming packet, depending on the circumstances."

"*Daehak-ro and colleagues* [3] High-performance computing (HPC) depends on network security, particularly when supercomputing services are provided across public networks. As supercomputer operators, we put in place a variety of security solutions, such as intrusion prevention systems (IPSs), firewalls, web application firewalls, and anti-DDoS software, to ensure the safe use of supercomputing resources. Potential risks are added to the firewall rules for access restriction based on specified security policies once abnormal activity is detected by anti-DDoS, IPS, and system access logs. After analysing the status change patterns for rule policies created as a result of human mistake among these new firewall log events, 289,320 data points were extracted over a four-year period."

"*Malak Aljabri et al.* [4] We are currently experiencing unprecedented network security issues. In actuality, this lends credence to the growing importance of network security. Firewall logs are important sources of data, but they are difficult to analyse. Artificial intelligence (AI), machine learning (ML), and deep learning (DL) have gained popularity as methods for developing robust security measures because of their ability to swiftly handle complex threats."

Iqbal H. Sarker [5] In the current Fourth Industrial Revolution (4IR or Industry 4.0), data from the Internet of Things (IoT), cybersecurity, mobile, social media, business, and health are all widely available in the digital world. Comprehending artificial intelligence (AI), specifically machine learning (ML), is crucial for conducting intelligent data analysis and developing the corresponding automated and intelligent applications. The field of machine learning encompasses a wide variety of algorithms, such as reinforcement learning, supervised, unsupervised, and semi-supervised. Furthermore, deep learning, which is a subset of a broader family of machine learning techniques, can analyse vast amounts of data intelligently. In this work, we present a comprehensive study of different machine learning methods that can be applied to enhance the intelligence and capabilities of an application.

"*Subrato Bharati and Prajoy Podder* [6] The Internet of Things (IoT) connects a number of smart devices that can talk to each other with minimal human involvement. In computer science, the Internet of Things is evolving rapidly. However, there were additional security challenges because of the IoT systems' cross-cutting nature and the variety of components used to implement such schemes. Fundamental security issues arise from the inefficient implementation of application security, authentication, encryption, and access network protocols in IoT devices. By strengthening existing security procedures, the IoT environment may be effectively safeguarded. In recent years, machine learning (ML) and deep learning (DL) have made significant strides in several important applications. Therefore, DL/ML techniques are required to change the protection of IoT systems from just permitting safe communication between IoT systems to security intelligence systems. This review aims to integrate a comprehensive analysis of ML systems and the latest developments in DL approaches to further improve IoT device safety methods. Nonetheless, several recent advancements in machine learning and deep learning for IoT security demonstrate how it may facilitate additional research."

"The review of existing literature on the application of machine learning algorithms in cybersecurity and market segmentation reveals a growing convergence of these domains, particularly within the context of big data. Farah Khoshaba et al. [1] and Qasem Abu Al-Haija [2], Machine learning has demonstrated significant potential in enhancing both data security and market insights, leveraging techniques like supervised, unsupervised, and reinforcement learning to effectively address specific challenges. Daehak-ro et al. [3] and Malak Aljabri et al. [4] In cybersecurity, the use of machine learning as a firewall decision-making tool has been shown to provide proactive threat detection and anomaly identification, significantly improving the ability of systems to defend against sophisticated cyber-attacks. Supervised learning models, which utilize historical data, are particularly effective for detecting patterns indicative of malicious activities. However, challenges such as false positives, data imbalance, and evolving threat landscapes require continuous model updates and hybrid approaches to optimize performance. Iqbal H. Sarker [5] and Subrato Bharati and Prajoy Podder [6], In the realm of market segmentation, machine learning techniques such as clustering (an unsupervised learning method) are increasingly used to identify distinct consumer groups and adapt to changing consumer behaviors. Reinforcement learning, with its focus on continuous improvement, has also been highlighted as a powerful approach to making market segmentation more responsive to dynamic consumer preferences. Nevertheless, issues related to data privacy, the heterogeneity of big data, and computational complexity are ongoing concerns that need to be addressed. The literature underscores the importance of integrating machine learning solutions for both data security and strategic business operations. However, most studies tend to address these domains in isolation, leading to missed opportunities for synergistic benefits. Combining machine learning for cybersecurity with market segmentation,

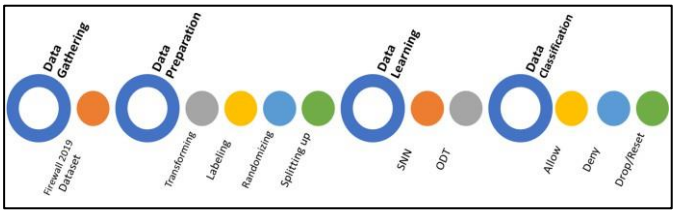
as proposed in this research, aims to bridge this gap and demonstrate the mutual advantages of a unified approach."

PROBLEM STATEMENT

The problem addressed is the development of a dual-function machine learning algorithm that serves as both a firewall decision tool and reinforcement for market segmentation, tackling challenges in cybersecurity and big data utilization. This unified solution aims for proactive threat detection and dynamic, adaptive customer segmentation. [7,8,9,10]

"REGARDING DISCOVERY AND DECLARATIVE PROCESS MODELLING"

In order to improve the security and defence of communication networks, we want to provide a comprehensive machine learning-based framework that ensures an automated and intelligent decision-making process for the firewall system. Fig. 1 displays the flowchart diagram for the system development approach, which illustrates the systematic steps for the proposed system from the initial study and data gathering stage to the final conclusion stage, the categorisation stage. The graphic illustrates the four modules that comprise system development: data gathering, data preparation, data learning, and data classification. The modules will be discussed in the following subsections after Fig. 1.



"Fig. 1 Architectural diagram of the proposed classification framework"

"Data Gathering Module"

"Data is a crucial part of every intelligent system since it allows stakeholders and systems to base their judgements on verifiable facts and records. Data (numerical, category, image, etc.) are frequently collected into neatly arranged entries in a systematic dataset [11]. Problem conceptualisation, research enquiries, and the validation of hypotheses and findings can all benefit from dataset analysis. Since the goal of this study is to automatically and intelligently identify the security activities that firewall devices perform on network traffic, we have employed a dataset [12]."

"In order to automatically predict the actions that the firewall would do in response to network traffic data, a recently developed dataset called IFW − 2019 [12] was constructed from internet traffic records on firewall devices at a university (Firat University, Turkey). The output field of every sample record in the 65532 firewall log records that IFW 2019 gathers has four different categorisation labels: "allow," "deny," and "drop.""

"Data Preparation Module"

"Like any machine learning-based system, the dataset undergoes a number of pretreatment procedures to prepare it for use by the machine learning input layer for further processing and learning operations. This study is the result of processing the dataset we collected."

"Dataset Transformation"

"Since the dataset records are available as a.csv file with several rows and columns (separated by commas), where the rows represent the data samples and the columns indicate the characteristics, the dataset must be modified

using the "reset-both" function of the MATLAB system (our development). The data distribution across the different file classes is shown in Table 1."

"TABLE I: Statistics of Traffic Distribution of IFW-2019 [32]"

Actions	Allow	Deny	Drop	Reset-both
No. of Record	37640	14987	12851	54
Description	Permit the date packet	Block the data packet	Drop the data packet	Send TCP reset to both the client and server devices

"IFW − 2019 records are created using firewall log files with 11 features and one class label. Features are carefully selected as numerical datatypes to ensure that machine learning techniques are implemented effectively: source port, destination port, NAT source port, NAT destination port, elapsed time for flow (in seconds), total bytes, bytes sent, bytes received, total packets, packets sent, and packets received [13]."

The IFW-2019 dataset is really nominated for evaluation in this study since it is publicly available as a.CSV filetype and has a respectable number of unique samples that protect the classifier from the impact of a more common class. Furthermore, this dataset covers all common security measures for servers and firewalls on network traffic. Furthermore, it may be efficiently preprocessed and set up to generate multi- class categorisation for the firewall activities of the communication networks. Finally, IFW − 2019 can be updated, enlarged, stimulated, and changed into a double matrix to facilitate any additional machine learning or computation processes. At this stage, the dataset was also transformed into a matrix of features with corresponding samples (11 x 65532) and a vector of labels (1 x 65532).

"Dataset Labelling: Because the dataset class features are kept as categorical datatypes, these datatypes must be translated into numerical labels (labelling) in order for machine learning algorithms and computations to process them mathematically. In order to properly mark the target classes, we have used the one hot encoding approaches [17] as follows: Allow (100), Deny (010) and Drop/Reset (001)."

"Dataset Randomisation: By assuring randomised dataset samples, this step helps to improve the validation and testing phases by redistributing the dataset samples in a way that avoids any classification preferences. The shuffling technique, which shuffles the dataset's data samples through random places, is the data randomisation policy we utilised to achieve this."

"Dataset Splitting Up: The purpose of this step is to separate the data into three categories: testing, validation, and training. The $DivideRand$ technique has been employed as a dataset distribution policy that D. Data Classification Module."

"We have employed the SoftMax activation function (multi-class classifier) to compute the probability for the output classes. A vector of K real numbers ($\mathbb{R}$k) is normalised using the normalised exponential formula SoftMax to produce a probability distribution of K real number-probabilities ($\mathbb{R}$k) proportionate to the exponentials of the input values [14]. The final neurone output from the preceding layer, which is activated using the Sigmoid function σ (net), is first taken into consideration in order to compute the numerical probabilities for each class uses random indices to divide the targets into three sets. Consequently, the distribution of the dataset The percentages are as follows: 70% for training, 15% for validation, and 15% for testing. The dataset distribution figures are:"

net [1] = (∑ W1j · Ij) j=1

===⇒
$then$

"Training: 45,872, Validating: 9,830, Testing: 9,830"
$H = (ne^{[1]}) = 1_{\ 1+\ e^{-}n[1]}$

$for\ i$ = 1, 2, 3, … , n (1)

B. Data Learning Module
"In order to train and categorise the communication traffic records supplied by the IFW-2019 dataset into three classes, we created an inference system in this study utilising an SNN and ODT: Permit"

"Eventually, the output layer computations via SoftMax σ:" ℝk⟼ ℝk is defined as:

n

"Deny Drop/Reset. Since "reset-both" has a limited number of samples (i.e., only 54 samples), we have consolidated both "reset-both" and "drop" operations into a single class in the third class."

net [2] = (∑ V1j · Hj) j=1

===⇒
$then$

1. "*Shallow Neural Network (SNN*): In SNN, data introduced [2]"

en[2]i

"to the network goes through a single hidden layer of pattern"
$O = \sigma(net$

) =
i K
j=1
en[2] i
$for\ i$ = 1, 2, 3, … , K (2)

acknowledgement. In our SNN, the input vector (I) has 11 inputs (I1; I2; I3; … ; I11) that connect the 11 characteristics of IFW − 2109 to the 150 neurones at the hidden layer (H1; H2; H3; … ; H150) in a completely connected manner. Additionally, all neurones are fully coupled to the output layer's three neurones (A1; B2; and C3), generating $Softmax$ probabilities for the respective classes (Class_1="allow"; Class_2="deny"; Class_3="drop/reset" ;). Lastly, W and V, from left to right, are the trainable weight vectors that correspond to each of the parametrised layers. Furthermore, we examine a neurone unit with a single output (H) and an input vector (I) of n components in order to illustrate the symbolic representation for the individual neurones. All of

the input vector *I*'s elements are multiplied by the appropriate weights in the weight vector W for each neurone, and the dot product of the weighs and inputs (W*T*) is then delivered to the intersection of the summation process.

"Ultimately, the $\mathbf{f}et$ value is formed by adding the bias *b* to the dot-product."
2) "*Optimised Decision Tree (ODT):* Effective and widely used tools for prediction and classification are decision trees. In knowledge systems like databases, decision trees serve as human-understandable rules. We have set up the tree in our ODT with one response variable (the goal variable) and eleven predictors. With a maximum of 30 splits using 30 iterations and 5-fold cross- validation, the tree split also adhered to the split requirement of maximum deviation reduction."

"Table 2 shows an example of the output from SoftMax classification. The classifier will always choose the label with the highest probability value for each instance based on the numerical probabilities given in the table."

"TABLE II: SAMPLE OUTPUT FROM SOFTMAX CLASSIFICATION"

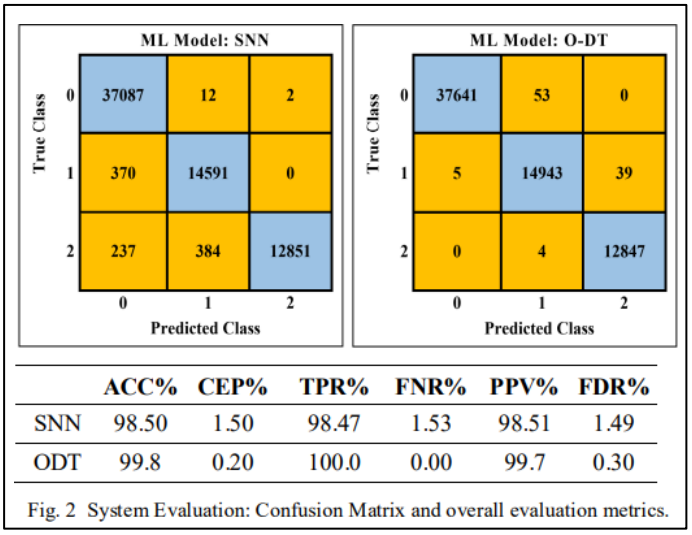
"Class label"	**"C1"**	**"C2"**	**"C3"**
"Actions"	"Allow"	"Deny"	"Drop/Reset"

"RESULTS AND DISCUSSION"
"The IFW − 2019 dataset was used for the training and testing stages in order to create and assess the suggested IFW classification system. The predictive model is designed to distinguish between three classes of network packets: "allow," "deny," and "drop/reset." MATLAB 2020b is used to implement the suggested predictive model on a standard laptop. Additionally, MEX computation (MATLAB executable) has been employed for gradient calculations and neural network training and simulation in order to maximise memory and training speed. Additionally, the original dataset was preprocessed before being used with machine learning algorithms. The preprocessing module is in charge of transforming the IFW − 2019 raw traffic data into a matrix of labelled features that the classification system's supervised learning component can use to train. In conclusion, Table III displays the test-bench environment's setups and characteristics."

"TABLE III: SUMMARY OF SYSTEM DEVELOPMENT ENVIRONMENT"

Specification	**Description**
Computing platform	High performance commodity PC with: • Intel I7-8550U CPU • 4.00 GB GPU NVIDIA GF 940MX
Supervised learning techniques	W-SNN with 150-Hidden Neurons using Three-Neuron Output Layer. ODT with 30 Splits using Maximum Deviance Reduction.
Optimisation Technique	Scaled conjugate gradient backprop. [21] for SNN Bayesian Gradient Optimisation [22] for ODT
NN Performance Analysis	Cross-Entropy Loss (LCE) Function [23] for SNN Mean Squared Error (MSE) Function [24] for ODT
Classification Learner	Linear learning algorithm Initial Learning Rate ($\alpha = 0.0001$)
Validation Frequency	6-Fold Cross Validation/Randomly Performed at Every Run.
Number of Epochs/Iterations	287 Epochs for SNN 30 Iterations for ODT

"The multi-class confusion matrix {true positives (\)), true negatives (\^), false positives ($]), false negatives ($^)}, classification accuracy (_``), classification error percent (`a]), true positive rate (TPR), false-negative rate (FNR), positive predictive value (PPV), and false discovery rate (FDR) are some of the evaluation metrics we used to assess the developed classification model in order to gauge the system performance [34]. The confusion matrix findings for our three-class classifier are thus displayed in Fig. 2. We have calculated the aforementioned assessment metrics for our classification model based on the values obtained for ⟨TP, TN, FP, and FN⟩."



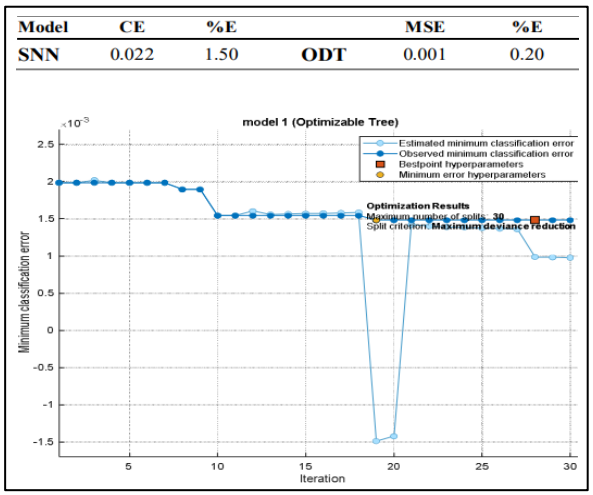
	ACC%	CEP%	TPR%	FNR%	PPV%	FDR%
SNN	98.50	1.50	98.47	1.53	98.51	1.49
ODT	99.8	0.20	100.0	0.00	99.7	0.30

Fig. 2 System Evaluation: Confusion Matrix and overall evaluation metrics.

"Additionally, because the classification model's goal is to generate output values that are as close to the true values as possible, the model's trainable weights are iteratively adjusted with the goal of minimising the Cross-Entropy Mean Squared Error (g a) value for the ODT model and the Cross-Entropy Loss (def) value for the SNN model. In order to penalise the probability according to its distance from the true value, the loss (def * P g) is computed by comparing the SoftMax probability (hS) for each predicted class to the true class label 5+S 9 [35]. The LCE or MSE loss of a perfect model is zero. Figure 3 (A) displays the mean secured error vs. iteration number plot indicating the optimal point hyperparameters for the ODT model, and Figure 3 (B) displays the cross-entropy for training, validation, testing, and best curves. Additionally, the findings for the SNN model's Cross-Entropy (CE) and Percent Error (%E) and the ODT model's Mean Squared Error (MSE) and Percent Error (%E) are shown in Table IV. Minimising CE or MSE in either model leads to good classification; zero indicates no error, and lower values are preferable. The percentage of samples that are incorrectly classified is shown by the percent error for %E. Whereas 100 denotes the greatest number of misclassifications, 0 denotes no misclassifications."

"LOSS AND ERROR VALUES FOR BOTH MODELS (SNN/ODT)"

Model	CE	%E		MSE	%E
SNN	0.022	1.50	ODT	0.001	0.20



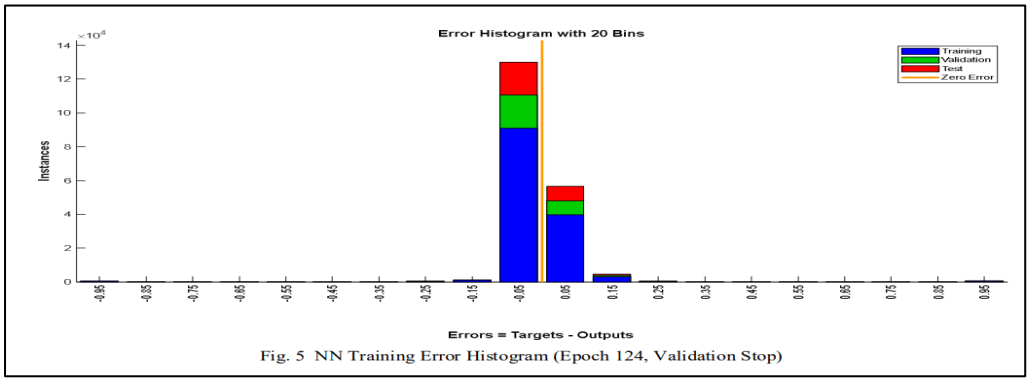


Fig. 5 NN Training Error Histogram (Epoch 124, Validation Stop)

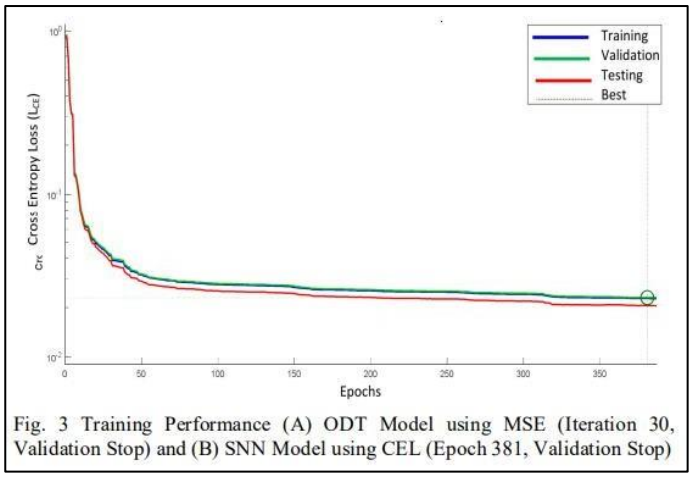


Fig. 3 Training Performance (A) ODT Model using MSE (Iteration 30, Validation Stop) and (B) SNN Model using CEL (Epoch 381, Validation Stop)

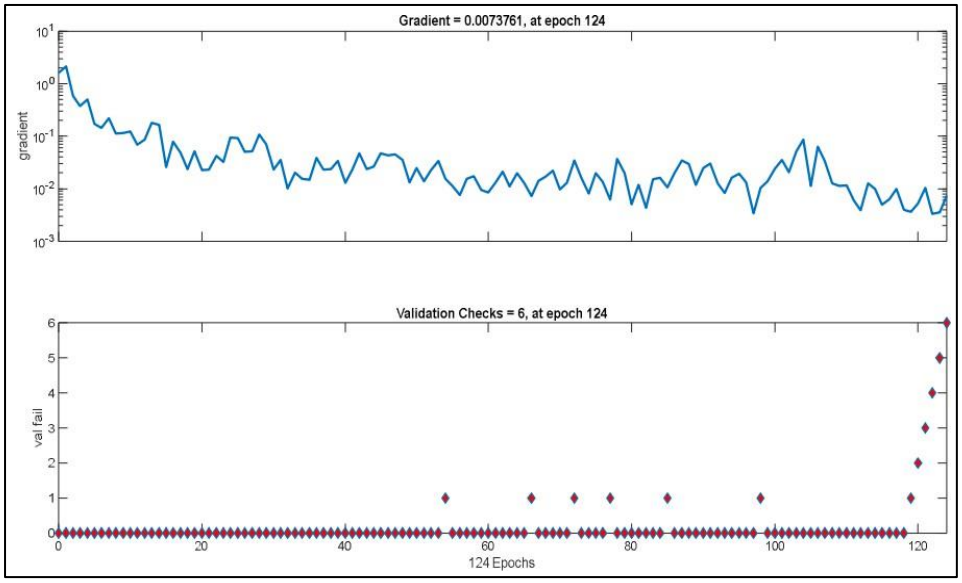
"Additionally, we have examined our 3-class classifier's receiver operating characteristic (ROC) curve. The ROC curve shows the relative trade-offs for different threshold settings between the false positive rate (costs) at the x-axis and the genuine positive rate (benefits) at the y-axis. When a continuous random variable (5i) is compared to a predetermined threshold (j), the classification model is often formed. As a result, the instance is categorised as "positive" if i > j and "negative" otherwise. Consequently, it is possible to compute the true positive rate (\]) and false positive rate ($]) completely for a given threshold (j). Thus, using \ as the variable parameter,
)` lPplots parametrically \]5\9 vs $]5\9. Figure 4 shows the ROC curves for our three- class classifier for the SNN and ODT models. The classifier nearly offers a flawless classification instance, recording 99.0% and 100.0% for

the area under the curve (AUS) of the SNN model and ODT model, respectively, since nearly all experiments produce a point in the upper left corner (0,1) of the ROC space. A histogram of MSE errors for the training, validation, and testing datasets is also displayed in Fig. 5. Twenty bins have been created out of the whole residual range. It is evident from the figure that the majority of MSE values are getting closer to zero. Additionally, the quality of the suggested machine learning model is reflected in the histogram error bars, which appear to follow a normal distribution curve. Furthermore, as the training dataset contains the majority of the dataset records within the employed dataset (i.e., 70% of the data samples belong to the training dataset, 15% belong to the validation dataset, and 15% belong to the testing dataset), the majority of counted mistakes pertain to the training dataset. Colour conventions are used to show these error scenarios, with blue denoting the training error residuals, green denoting the validation error residuals, and red Additionally, this figure shows the training's current state or progress at a certain point in time while it is ongoing. In our instance, the yellow line denotes the zero-error value, and the six validation errors that are indicated relate to the training error residuals. It indicates that when the six validation check faults are generated at the same time, the training will end. As is evident, the validation process was terminated after 124 epochs, during which the model initially encountered six validation check faults from the start of the training procedure."

"Lastly, we benchmarked the IFW classification system by contrasting its performance with other cutting-edge machine-learning-based firewall-action classification systems in terms of the classification accuracy measure in order to better understand the benefits of the suggested approach. The following lists the comparisons."

"As a result, it can be seen that the suggested IFW model outperforms other relevant machine learning-based models in terms of classification accuracy for the IFW 2019 dataset by an improvement percent (IM%) of ≈ (2.7% – 30.5%). The improvement percentage is determined by dividing the accuracy improvement of our model by the accuracy of the current model in the manner described below:"

$$IM = \left[\left(\frac{\textit{Our Model Accuracy}}{\textit{Other Model Accuracy}} \times 100\right) - 100\right]\%$$



"CONCLUSION"

"This study presents a machine learning algorithm designed to address dual challenges faced by organizations: safeguarding data against evolving cyber threats and effectively utilizing big data for market segmentation. The proposed algorithm integrates both a firewall decision-making system and a reinforcement mechanism for dynamic market segmentation, providing a comprehensive solution to the unique challenges in each domain. By leveraging supervised learning techniques for cybersecurity, the model is capable of detecting threats in real-time, adapting continuously to new patterns of malicious activity. Simultaneously, the application of unsupervised and reinforcement learning in market segmentation allows for the dynamic adaptation to changing customer behaviors, enhancing the precision and relevance of targeted marketing strategies. The integrated approach offers a significant advantage over traditional siloed methods by simultaneously improving data security and extracting actionable insights. The findings demonstrate that a unified machine learning framework can enhance the accuracy of threat detection while providing meaningful, real-time segmentation in a scalable manner. The conclusion drawn from this work is that machine learning can be effectively harnessed to address two critical, often separately handled aspects of big data: security and market insights. Future research could focus on refining the scalability of such a model, addressing computational complexity, and exploring deeper integration with real-world enterprise environments to maximize the benefits for cybersecurity and marketing simultaneously."

REFERENCES

1. Authorized licensed use limited to: b-on: UNIVERSIDADE NOVA DE LISBOA. Downloaded on July 02,2022 at 11:34:28 UTC from IEEE Xplore. Restrictions apply.

2. Qasem Abu Al-Haijaa,*, Abdelraouf Ishtaiwia, a Department of Data Science & Artificial Intelligence, University of Petra, Amman 1196, Jordan, Corresponding author: , qasem.abualhaija@uop.edu.jo

3. National Supercomputing Center, Korea Institute of Science and Technology Information, 245 Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea, Author to whom correspondence should be addressed. Appl. Sci. 2024, 14(11), 4373; https://doi.org/10.3390/app14114373 Submission received: 25 April 2024 /
Revised: 16 May 2024 / Accepted: 20 May 2024 / Published: 22 May 2024

4. Aljabri, M.; Alahmadi, A.A.; Mohammad, R.M.A.; Aboulnour, M.; Alomari, D.M.; Almotiri, S.H. Classification of Firewall Log Data Using Multiclass Machine Learning, Models. Electronics 2022, 11, 1851. https://doi.org/10.3390/ electronics11121851

5. Sarker IH. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Comput Sci. 2021;2(3):160. doi: 10.1007/s42979-021-00592-x. Epub 2021 Mar 22. PMID: 33778771; PMCID: PMC7983091.

6. 1,2Institute of Information and Communication Technology (IICT), Bangladesh University of Engineering and Technology (BUET), Dhaka-1205, Bangladesh Correspondence should be addressed to Subrato Bharati; subratobharati1@gmail.com

7. Moustafa, N., Slay, J., & Creech, G. (2019). Big data analytics for intrusion detection system: Statistical decision-making using machine learning techniques. Journal of Big Data, 6(1), 1-22.

8. Sahu, M., & Saini, M. (2020). A study on machine learning approaches in intrusion detection system. Journal of Information Security and Applications, 55, 102620.

9. Nunan, D., & Di Domenico, M. (2019). Big data: A normal accident waiting to happen? Journal of Business Research, 97, 222-231.

10. Kumar, V., Anand, A., & Gupta, S. (2021). Reinforcement learning in marketing: Framework, applications, and research agenda. Journal of Business Research, 123, 297-308.

11. Q. A. Al-Haija, S. Zein-Sabatto, "An Efficient Deep-Learning-Based Detection and Classification

System for Cyber-Attacks in IoT Communication Networks" Electronics, MDPI, vol. 9, no. 12: paper no.2152., 2020.

12. F. Ertam, M. Kaya, "Classification of firewall log files with multi-class support vector machine," in Proc. Of 6thInternational Symposium on Digital Forensic and Security (ISDFS), Antalya, pp. 1-4, 2019.

13. UCI: Machine Learning Repository, “Internet Firewall Data Set”, Center for Machine Learning and Intelligent Systems, 2019.

14. Fei-Fei. CS231n: Convolutional Neural Networks for Visual Recognition. Computer Science, Stanford University. Available online: http://cs231n.stanford.edu, 2019.

15. J.J. Praise, R.J Raj, J.V. Benifa, “Development of Reinforcement Learning and Pattern Matching (RLPM) Based Firewall for Secured Cloud Infrastructure”, Wireless Personal Communication, Springer, vol.115, p.p. 993–1018, 2020

16. G. Bendiab, S. Shiaeles, A. Alruban, N. Kolokotronis, “IoT

17. Malware Network Traffic Classification using Visual Representation and Deep Learning”, in Proc. Of 6th IEEE Conference on Network Softwarization (NetSoft), Ghent, Belgium, 29 June–3 July 2020; pp. 444–449

18. K.E. Koech, “Cross-Entropy Loss Function”, Medium: towards data science, 2020.

19. D. Appelt, C. D. Nguyen, A. Panichella, L. C. Briand, "A MachineLearning- Driven Evolutionary Approach for Testing Web Application Firewalls," IEEE Transactions on Reliability, vol. 67, no. 3, pp. 733-757, 2018, doi: 10.1109/TR.2018.2805763.

20. R. Shire, S. Shiaeles, K. Bendiab, B. Ghita, N. Kolokotronis, “Malware Squid: A Novel IoT Malware Traffic Analysis Framework Using Convolutional Neural Network and Binary Visualization”, in Proc. Of Internet of Things, Smart Spaces, and Next Generation Networks and Systems. Lecture Notes in Computer Science; Springer, vol.11660, 2019.

21. Baptista, S. Shiaeles, N. Kolokotronis, “A Novel Malware Detection System Based On Machine Learning and Binary Visualization”, in Proc. Of IEEE International Conference on Communications (IEEE ICC), China, pp. 1–6, 2019.

22. R. Garg, “Types of Classification Algorithms”, Analytics India Magazine, 2018 F. Ertam, M. Kaya, "Classification of firewall log files with multi-class support vector machine," in Proc. Of 6thInternational Symposium on Digital Forensic and Security (ISDFS), Antalya, pp. 1-4, 2019.