[1] **Mohammad Taghi Sadeghi**

[2] **Hiba Alzubaidi**

Strengthening Wireless Network Security: Supervised Machine Learning-Based Intrusion Detection for Enhanced Threat Mitigation

Journal of Electrical Systems

Abstract: - With the wireless networks deemed vital to our laced and unetched world, it has become fundamental to safeguard these networks against the various cyber threats that are on the rise. As a research proposal I show concern about the application of a Supervised Machine Learning-Based Intrusion Detection system, making use of Multilayer Perceptron (MLP), Support Vector Machine (SVM), and Logistic Regression (LR) algorithms. The survey begins with the introduction of an intrusive inter-dataset technology that merges two catalytic datasets into a single efficient model that can address both model security and data sensitivity concerns. The evaluation of several machine learning algorithms within the proposed framework is a core component of the analysis. The accuracy measure as well as others such as recall, precision, and F1 score, are applied to evaluate the algorithms' level in detecting intrusions from different datasets.The investigation shows that the intra-dataset routing noticeably improves the detecting modules' efficiency. Notably, the model of MLP algorithm has enhanced recall which in turn shows the enhancement of the positive instance identification. SVM shows increased precision, which accounts for the improved accuracy since correct names of positive cases are produced more often. LR finds it overall enhance the precisions thus attesting to its efficiency in correct deductions. The research documents the multifaceted nature of IDS, prompting the intervention of the machine learning algorithms with ddoswdelcome intrusion detection system. A study which provides a practical guide for network administrators and security professionals hinting on how to select algorithms to meet the distinct security needs in their enterprises. These findings contribute to the open discourse on wireless network security and serve as a base for future study of this engaging research topic in constantly developing field of intrusion detection.

Keywords: Wireless Networks; Intrusion Detection; Supervised Machine Learning; Multilayer Perceptron; Support Vector Machine; Logistic Regression.

I. INTRODUCTION

Today, the information flow is so versatile that we are connected to dozens of users, sometimes even on the other side of the planet; therefore, discernment of the security threats that scattered individuals may face is a critical issue. The advent of the WiFi devices and the rising complexity of the networks have escalated threats targeted at intrusion and unauthorized access to a high critical level. Conventional security measures often do not meet the standards in recognizing and catching up with intricate attacks, therefore the machinery use of learning is incepted by the successor to cater for intrusion detection. However, though a riskless intrusion expedition system conceived for the wireless networks of the respective supervised machine learning methods [1] is what the challenge entrenched in that is, it is indeed a challenge.

Primarily, wireless networks feature a specific set of hazards for detecting intrusion because these networks are prone to eavesdropping, jamming, and spoofing attacks which are due to the open nature of wireless communication. Factors that influence the signal quality such as varying intensity, interference, and mobility are just as challenging, because identification of malicious and abnormal behavior as legitimate system activity can be very difficult. It is then necessary to work out a weapon which will be able to store wireless networks' subtleties rightly [2].

Secondly, during acquirement labeled datasets for calibration of supervised machine learning techniques is the serious issue. The hard accurate endorse and naming of real world wireless network’s traffic including both usual and intrusion happenings are incredibly time and resource consuming. The evolving nature of the attacks requires, regularly, updates to the dataset for incorporation of the emerging IO patterns which on the other hand, poses a challenging constraint in achieving accurate and comprehensive representation of different IOs [3].

Thus, after that, it is crucial to consider waveform, timing, and payload during intrusion detection in wireless networks. Feature that are applicable to wired networks may not hold any relevance in the wireless side. The discovery of occurring traits which represent the specificities of wireless networks mode, such as signal strength, packet loss, and channel usage is of paramount importance. Furthermore, applying the feature selection techniques

[1] Faculty of Technical and Engineering, University of Qom, Qom, Iran

[2] Technical College Of Management/Baghdad, Middle Technical University, Baghdad, Iraq

mt.sadeghi@srbiau.ac.ir, queen.of.cipher@gmail.com

Copyright © JES 2024 on-line : journal.esrgroups.org

is essential for managing high-dimensional data, and it will immensely improve the process successfully and effectively. Following Kenman et al. [4].

And, finally, making diagnostic scrutiny of the performance of trained models and picking the most fitting algorithm for the unsupervised context of the deployment allows for the identification of the most significant challenge. Assessment should be comparable to requirements of wireless networks, insisting on something similar to noise reduction to help avoid hijacking of relevant data and messages. Therefore, accomplishing complicated assessment models tailored to the wireless network characteristics and specifications is essential [5].

Secondly, installing an intrusion detection system in such against wireless networks, needs highly detailed attention. The system has to deal with ever-growing wireless traffic with a high-speed, high-volume nature and the demand for a quick reaction to the traffic jams on the roads. Scalabi'ly and adaptibility are the mode of the network to adjust on the changing network conditions and the shaoght, new type attacks. While the integration of the wireless into the network architecture is less of a problem, the challenges brought forth such as underperformance bottlenecks, or perhaps interference of the functionality are still on the table. Consequently, dealing with these difficulties and building a good intrusion detection system using supervised machine learning techniques, particularly designed for wireless networks, is vital. It serves this purpose: to offer security, comprehensiveness, and accuracy, which are fundamental in the context of safeguarding of wireless networks and blocking any attempted intrusions. "Is it possible to apply guided machine learning algorithms in intrusion detection for wireless networks in view of the peculiar attributes of wireless networks, available labelled datasets, choice of correct features, measurement of model performance, and putting into operation of such models in real time wireless network systems?".

II. Research Literature Review

A. Wireless Communication

This is because the wireless technologies has become a major indicator, pointing out, the transformation the way people communicate, process information and use digital platforms among cognoscente. In contrast to the traditional mode of communication the latter does not even consider physical cables and rather employs RF data signals. The wireless networks invention has given start to the new era of universal connectivity and mobility in which communication among a large variety of devices and services will be without limitations. Over and above this powerful shift, this offers freedom from being bound by physical cables, allowing for a whole new world in the digital environment to be experienced with unrivaled levels of flexibility and access. The media where it is wireless is one of the significant aspects of the mobile devices, smart appliances, and interconnected systems. It is the fashion for the modern people have to do their communication by using the wireless networks that provide a universal and interconnected infrastructure that surpasses the conventional boundaries. The emergence of this paradigm emphasizes the evolution of the life of people via wireless technology at the expense of their social intercourse, working, and mobility [7].

B. Intrusion Detection

In the modern and integrated digital environment, where computers and network connection are more and more common, it is imperative that cyber security protection is not only more critical that it has never before been. Namely, intrusion detection, which functions as the alert police, ensures the safety of our networks and systems. This opening paragraph discusses a topic of Intrusion Detection that comprises of main principles and helps the users to be aware of the importance of it for network security and role that it plays to protect the digital assets [8]

C. Feature Extraction

Feature extraction represents a very significant element of both data analysis and machine learning in general, and needless to say, its relevance to network security is emphasized. This procedure is the filtering, slicing and combing process where the original data gets reduced into a collection of kind and converting values which make up the selected features or variables which are important. Feature extraction plays a huge role in representations of network info to be used in analysis and intrusion detection among network security processes. Through the reduction of the abundant and mostly redundant network information into a set of key features that are crucial for detection, the identification and tracking of malicious activity improve [9].

III. RESEARCH METHODOLOGY

A. The proposed framework

The study is based on an inter-dataset model for syndication of datasets, aiming to target the weaknesses and building on strengths of two interconnected sensors. A relational basis of datasets importantly introduces an all-new aspect to the evaluation procedure by creating the groundwork of a holistic realization of system performance in varying situations. The details of this strategy are explained in Figure 10 which is the visuluiar aide of this proposed method.

Dataset Interconnection: A given research's datasets are intrinsically connected, with one being the subset of the other. This relationship lets the two datasets share a seamless data set testing model, which the machine learning models use the provided dataset to test its accuracy. This interconnectedness not only provides a thorough appraisal of model convergence and generalization in distinct datasets but also facilitates extrapolation of the models to solve problems.

Evaluation Workflow: Datasets go through pre-processing, and feature selection is among key procedures during it. Using that, each data set is divided into trains and test groups. After the subsequent step is multiple instances of the model that possess different hyperparameters each being trained on the training set. Once the training set is done with, that will be followed with the test set to determine which model will perform better by looking into the characteristics of this set. Also, it follows the another common strategy used with the in-dataset process, which improves the consistency and comparability.

Cross-Dataset Strategy: We propose to use an unconventional approach for the evaluation in the research. The cross-dataset strategy is used in place of traditional evaluation methods. Using the second dataset rather than the test set of the same dataset is the implementation of feature removal and model training on a single data. This is followed by the actual assessment of the model, which is done on the test set of the second data source. This cornucopia strategy indeed offers a more complex analysis of the model's robustness and flexibleness through various data domains are interrelated.

Dual Execution of Cross-Dataset Strategy: Research uses two data sets one by one, so that cross-dataset strategy is completed the twice. The initial pass of the algorithm / training on dataset A and testing on dataset B, is followed by a second iteration with dataset B training and dataset A testing. These cohort approaches are not only computationally fast and therefore extremely robust, but also reduce the concerns associated with sensitive information within the data. Rather than operating on solely one dataset, the study serves as a tool for better comprehending the intrusion detection model performance especially when applied to different datasets.

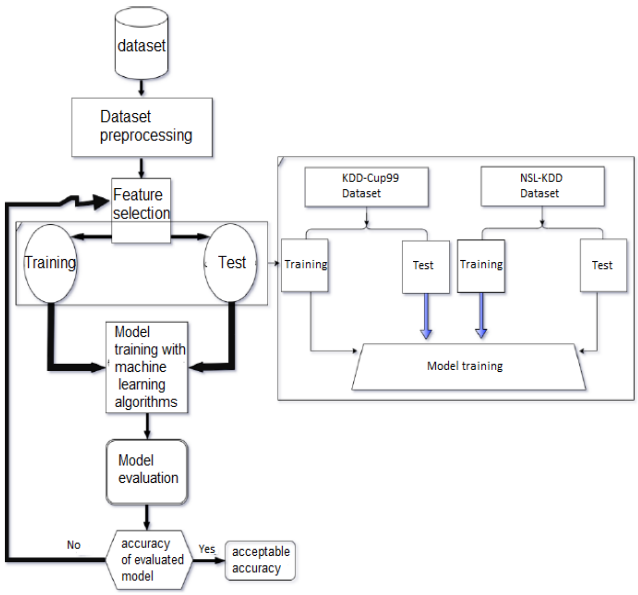

Figure 1. Steps of the proposed method

B. Multilayer perceptron

Multilayer Perceptron (MLP) is a variety of Artificial Neural Network (ANN) which is defined by a large number of neurons that are attached to each other through the process of stochastic gradient descent. In which case the data flows also directionally feed-forward as a “neural network” inside the network, without forming terrorism cycles.

Multilayer Perceptron (MLP) represents one of types of artificial neural networks, which possesses the ability to connect several layers of nodes, each playing role of a neuron. It has also stacked by input layer, hidden layer and output layer that at least one. Also a neuron of a particular layer is connected to every neuron of the following layer, causing the construction of a network which is capable, in its turn, of teaching complex patterns and relations of data. The meaning of the word "multilayer" becomes apparent through its use, as it reflects what distinguishes MLPs from single-layer perceptrons: an addition of hidden layers.

So in an MLP, there is a weight that is connected with every neuron and a summation of that weight for each neuron is taken into account through an activation function. This gives the network ability to identify non-linear patterns under generic logistic regression function. The training is initiated by souresing between the output predicted by the network, and the actual target value. Such process of improvement of the performance is carried out sequentially and as a result it facilitates adaptation process.

Artificial neural networks (MLPs) are very often implemented in different areas such as image recognition, speech or natural language processing and financial forecasting [10].

C. Support vector machine (SVM)

SVM, is a supervised machine learning algorithm that employs a support vector machine for classification and regression tasks. Definition lies in the fact that the task is just to find a hyperplane in a high-dimensional space that will be the most effective in terms of point classification into distinct classes. While in the binary classification tasks, the hyperplane would be adjusted so as to make the margin maximum by considering the distance from the hyperplane to the closest data points on each class. SVM is mostly dealing with intricate data sets, which brings its applicability of the problems to other nonlinear types of problems if the kernel functions are utilized to map the inputs into higher-dimensional spaces and render decision boundaries linear.

The main concept of SVM is to find the support vectors (which are the data points near the decision boundary) and then come up with hyperplane equations. These support vectors are where the decision line pass which provide this model with the ability to generalize and act as the characteristic features among the samples. SVMs fall into category of algorithms that are known for their high level of robustness and effectiveness in very high dimensional space and as such are useful in tasks of image classification, text classification and several other applications. They on the contrary can become sensitive to a choice of a kernel function and numbers of parameters, which may, in fact, require some fine-tuning for the best efficiency [11].

D. Logistic Regression (LR)

LR is an advanced statistical algorithm as well as the most important one in the machine learning field and it is employed in tasks where a binary classification method is required. The name logistic regression is misleading since logistic regression is to be used for categorical data and not linearly related data. The foundation of logistic model is the ability to estimate the exact probability that a given signal bears or does not bear the certain classes. This is called logistic regression because it uses logistic function (the sigmoid function), which is polynomial in nature, to process the linear combination of features to give the probability between 0 and 1, where 0 means failure and 1 means success.

IV. DATA ANALYSIS

A. Data collection

1. KDD-Cup99 Dataset:

- We start off by employing in this strategy the KDD-Cup99 dataset as the first dataset. Since the original DIDE project in partnership with the US three-letter agency, DARPA and Cambridge, Massachusetts' Lincoln Laboratory at MIT, the KDD-Cup-99 dataset was constructed to test ID (intrusion detection) algorithms. The rich set of information obtained in the DIDE research project is the source code to set a quality control for the developed detection systems selection, i.e. the benchmark.

2. NLS-KDD Dataset:

- The second dataset employed in this research is the NLS-KDD data, which constitutes a carefully selected subset derived from the KDD-Cup99 dataset. [11] undertook a meticulous curation process for the NLS-KDD dataset, involving a comprehensive statistical analysis of the KDD Cup 99 data. This refined subset specifically addresses inherent issues within the original KDD Cup 99 data, presenting itself as a pertinent and improved dataset for the evaluation of intrusion detection systems.

By identifying and addressing limitations and anomalies in the original dataset, the NLS-KDD subset was crafted to enhance the quality and relevance of data for intrusion detection system evaluation. This subset, derived through a rigorous curation process, contributes to the reliability and effectiveness of the proposed inter-dataset strategy, offering a more robust foundation for the application of supervised machine learning in intrusion detection within the context of wireless networks.

Both datasets consist of 42 features or fields, encompassing 41 normal features related to network connections and a class feature. The class feature defines five distinct classes, including a normal class and four attack classes: DoS, U2R, R2L, and prob.

TABLE I. FOUR DATA-RELATED CLASSES

Class	Description
DoS	A group of attacks is said to target the component (accessibility) of information and thus prevent users from accessing the services provided in a network
R2L	In this type of attack, the attacker tries to gain control of the victim's machine remotely, using methods such as guessing users' passwords and buffer overflows. If this attack is carried out successfully, depending on the permissions and the level of access that the intruder has obtained, it can completely damage all three basic components of information security.
U2R	This type of attack is executed on the victim's machine and the attacker, who has the access level of a normal user, tries to take over the permissions of the root user (in Linux systems, the root user and in Windows systems, the user (Administrator).
Probing	In this category, which is also known as exploratory and detection attacks, the intruder uses various tools such as Nmap to scan the machines in the target network to gather the basic information needed to start the attack and also to find known vulnerabilities.

B. Methodology

The key premise of the cross-dataset evaluation strategy is to train a model on the first dataset (KDD-Cup99) and subsequently test it on the second related dataset (NLS-KDD). The goal is to assess the model's ability to generalize and maintain classification performance across different yet related datasets.

C. Preprocessing

The data preprocessing part is the most essential stage in the machine learning pipeline whose role is to tie the raw data with its learning algorithm in order to get the outputs. In this part, we will describe the preprocessing techniques that have been used on both datasets in order to turn the raw data into a reworked data that is easier, and more efficient to be worked on by machine learning. Figure 1 shows the most important stages of data preprocessing that include: removing duplicates, finding unneeded data having no relevant information, taking care about the presence of non-finite values, and scaling the data. This final purpose is the prevention of loss of information, meaning we get a high-quality and accurate power to both our feature selection and machine learning process.

D. Results

In this section, we present the results of the categorization technique employed on the KDD-Cup99 and NSL-KDD datasets, the improvement efforts on accuracy obtained using supervised algorithms per individual dataset as the focus. In Chapter III the participant observer goes into detail to explain the method of calculating these criteria.

Table 2 illustrates a data table with the summarized score for classification. The table plot includes a columns row to indicate training dataset followed by the test dataset on the inter-dataset evaluation. In particular, it enumerates the categorization effectiveness, training complexity, and the output complexity for all algorithm. Please have a look on Tables 1 and 2 for details of the results.

TABLE II. FOUR DATA-RELATED CLASSES

Algorithm	Dataset	1:Recall	2:Precition	3:F1	4:Accuracy
MLP	KDD-Cup99 , KDD-Cup99	0.997	0.997	0.997	0.998
	NSL-KDD , NSL-KD	0.998	0.987	0.990	0.981
	KDD-Cup99 , NSL-KDD	1.000	0.991	0.990	0.991
	NSL-KDD , KDD-Cup99	0.999	0.994	0.997	0.994

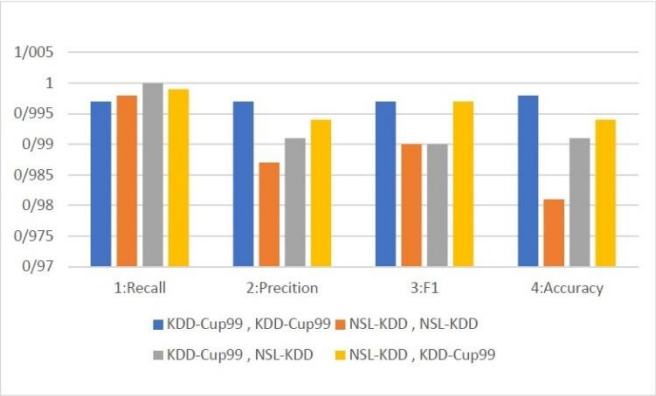

Figure 2. Results for the MLP algorithm

Table 2 in a nutshell offers a compact rating of how a well the MLP algorithm shows itself in respect to different datasets by means of quantitative classification performances on different scenarios. It become an important instrument that shows whether algorithm works completely or not and how well it can detect intrusion across various streams of data. The first row tells that in case, the model is trained and tested on the KDD-Cup99 database, it has an accuracy of 0.997 and also has more or less a good score in recollecting, precision and F1 scores.

TABLE III. RESULTS FOR THE SVM ALGORITHM

Algorithm	Dataset	1:Recall	2:Precition	3:F1	4:Accuracy
SVM	KDD-Cup99 , KDD-Cup99	1.000	0.995	0.998	0.995
	NSL-KDD , NSL-KD	1.000	0.984	0.992	0.983
	KDD-Cup99 , NSL-KDD	1.000	0.993	0.992	0.990
	NSL-KDD , KDD-Cup99	0.522	0.994	0.658	0.457

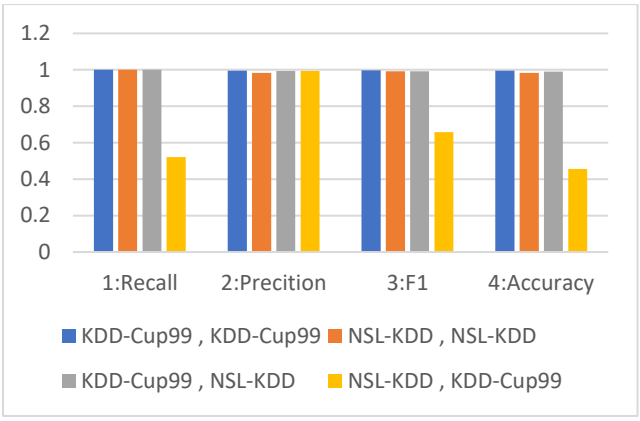

Figure 3. Results for the SVM algorithm

Table 3 provides a compact summary of how well the SVM algorithm performs on different dataset pairs, offering insights into its capabilities in intrusion detection across diverse data scenarios.

Each row in the table corresponds to the performance of the SVM algorithm on a specific dataset pair.

For example, the first row indicates that when the model is trained and tested on the KDD-Cup99 dataset, it achieves high recall, precision, F1 score, and accuracy.

The last row suggests that when the model is trained on NSL-KDD and tested on KDD-Cup99, the recall is lower, indicating challenges in identifying positive instances.

TABLE IV. RESULTS FOR THE LR ALGORITHM

Algorithm	Dataset	1:Recall	2:Precition	3:F1	4:Accuracy
LR	KDD-Cup99 , KDD-Cup99	0.997	0.996	0.954	0.997
	NSL-KDD , NSL-KD	0.977	0.974	0.857	0.983
	KDD-Cup99 , NSL-KDD	0.714	0.759	0.814	0.974
	NSL-KDD , KDD-Cup99	0.001	0.001	0.822	0.918

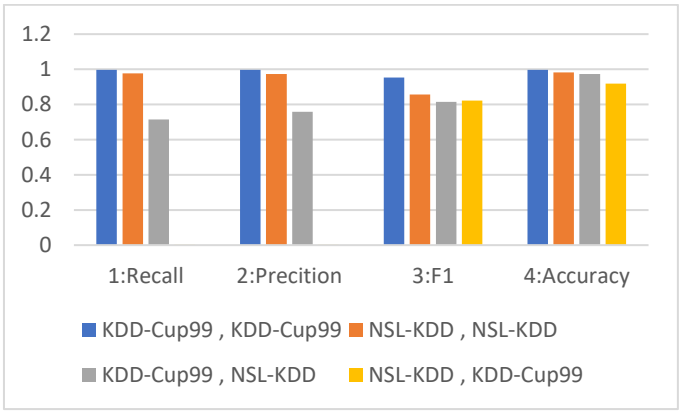

Figure 4. Results for the LR algorithm

V. CONCLUSION

The results associated with algorithm implementations indicate that intradataset evaluation in many cases serves as a reliable source of improving the performance of certain algorithms while at the same time aiding in variation in accuracy levels in a number of algorithms. Perssecution accuracy criterion is the main standard to be highlighted in this situation. As the results of this criterion, reproduced below and compared with those of the other algorithms, are reviewed, some compelling observations can be made.

A. Specifically

1. Multi-layer Perceptron (MLP) Algorithm:

- The evaluation indicates that the recall criterion within the MLP algorithm has experienced improvement. This suggests that the model's ability to correctly identify instances of the positive class, such as attacks, has been enhanced.

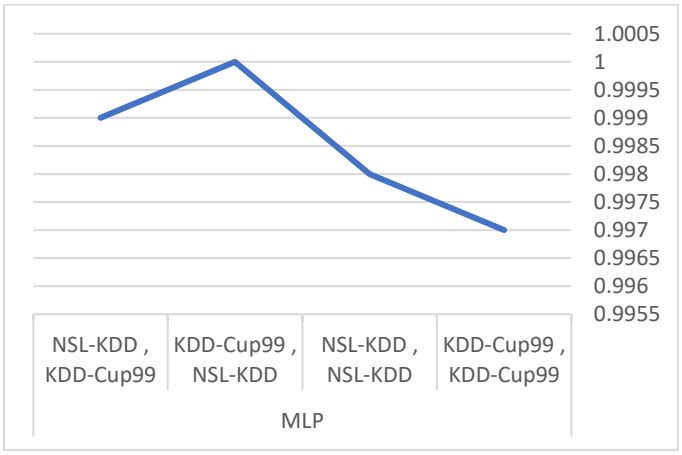

Figure 5. Recall for MLP algorithm

B. Support Vector Machine (SVM) Algorithm

For the SVM algorithm, the precision criterion has seen an increase. This signifies an improvement in the accuracy of positive predictions made by the model, showcasing its enhanced precision in identifying positive instances.

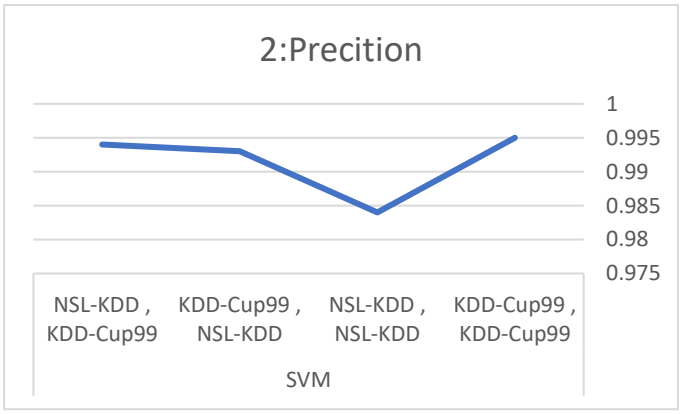

Figure 6. Precition for MLP algorithm

C. Logistic Regression (LR) Algorithm

Regarding the logistic regression and even the accuracy-improving attribute of the algorithm, there has been accuracy improvement. On the other hand, this means that the model will have the overall enhanced precision of the predicted data, resulting in a high level of accuracy.

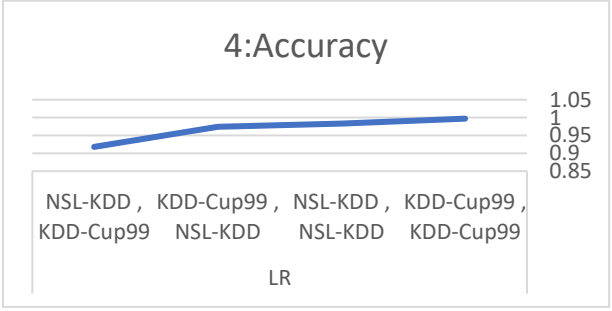

Figure 7. Accuracy for MLP algorithm

The often-cited adage that a man may only truly understand his home by traveling portrays a profound notion. An increase in recall score was observed with MLP leveraging, the SVM on precision and logistic regression in the overall accuracy. These ideas are in fact essential facts that result in a sophisticated comprehension of the strong and weak points of the technique as well as encourage the development of new algorithms to the sample data assessment under study.

REFERENCES

[1] S. Fraihat, S. Makhadmeh, M. Awad, M. Al-Betar, M. A., and A. Al-Redhaei, "Intrusion detection system for large-scale IoT NetFlow networks using machine learning with modified Arithmetic Optimization Algorithm," Internet of Things, pp.100-112, 2023.

[2] N. Mishra, and S. Mishra, "A Novel Intrusion Detection Techniques of the Computer Networks Using Machine Learning,". International Journal of Intelligent Systems and Applications in Engineering, vol (5s), pp. 247-260, 2023.

[3] B. S. Babu, G. A. Reddy, D. K. Goud, K. Naveen, and K. S. T. Reddy, "Network Intrusion Detection using Machine Learning Algorithms," In 2023 3rd International Conference on Smart Data Intelligence (ICSMDI) (pp. 367-371). IEE, (2023, March).

[4] D. Musleh, M. Alotaibi, F. Alhaidari, A. Rahman, and R. M. Mohammad. "Intrusion Detection System Using Feature Extraction with Machine Learning Algorithms in IoT," Journal of Sensor and Actuator Networks, 12(2), 2 , 2023.

[5] F. Naeem, M. Ali, and G. Kaddoum, "Federated-learning-empowered semi-supervised active learning framework for intrusion detection in ZSM," IEEE Communications Magazine, 61(2), pp. 88-94, 2023.

[6] R. Khandait, U. Chourasia, and P. Dixit, "Machine Learning Techniques in Intrusion Detection System: A Survey. In Computer Vision and Robotics," Proceedings of CVR 2022 (pp. 365-378). Singapore: Springer Nature Singapore, 2023.

[7] S. Lyu, L.Peng, and S. Y. Chang, "Investigating Large-Scale RIS-Assisted Wireless Communications Using GNN," IEEE Transactions on Consumer Electronic, 2024.

[8] M. Saied, S. Guirguis, and M. Madbouly, "Review of artificial intelligence for enhancing intrusion detection in the internet of things," Engineering Applications of Artificial Intelligence, 127, 107231. 2024.

[9] H. Tang, Y. Tang, Y. Su, W. Feng, B. Wang, P. Chen, and D. Zuo, "Feature extraction of multi-sensors for early bearing fault diagnosis using deep learning based on minimum unscented kalman filter," Engineering Applications of Artificial Intelligence, 127, 107138, 2024.

[10] S. Afzal, B. M. Ziapour, A. Shokri, H. Shakibi and B. Sobhani, "Building energy consumption prediction using multilayer perceptron neural network-assisted models comparison of different optimization algorithms," Energy, 282, 128446, 2023.

[11] A. Roy and S. Chakraborty, "Support vector machine in structural reliability analysis: A review," Reliability Engineering & System Safety, 10912, 2023.

[12] J. Ma, P. Dhiman, C. Qi, G. Bullock, M. van Smeden, R. D. Riley and G. S. Collins, "Poor handling of continuous predictors in clinical prediction models using logistic regression: a systematic review," Journal of Clinical Epidemiolog, 2023.