IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, May 2022
Impact Factor: 6.252

Prediction of Cyber-Attacks Using Data Science Techniques

Mr. Sudarsanam[1], Sudharsan M[2], Sakthivell V[3], Sakthi Vignesh P[4], Raghavendra Rao D[5]
Assistant Professor, Department of Cyber Security[1]
UG Scholor, Department of Computer Science and Engineering[2,3,4,5]
SRM Valliammai Engineering College, Chengalpattu, India

Abstract: *Cyber-attacks aim to destroy or maliciously manipulate a computing environment or infrastructure, as well as disrupt data integrity or crack all information. This poses a risk to the organisation, perhaps resulting in data loss. The data from device sensors is collected as big data, which has a wealth of information that can be utilised for targeted assaults. Although existing methodologies, models, and algorithms have given the foundation for cyber-attack predictions, new models and algorithms based on data representations other than task-specific techniques are required. Its non-linear information processing architecture, on the other hand, can be customised to learn alternative data representations of network traffic in order to classify different types of network attacks. In this study, we treat cyber-attack prediction as a classification issue, in which networking sectors must use machine learning approaches to forecast the type of network assault from a given dataset. The supervised machine learning technique (SMLT) is used to analyse a dataset in order to capture multiple pieces of information, such as variable identification, uni-variate analysis, bi-variate and multi-variate analysis, missing value treatments, and so on. A comparison of machine learning algorithms was conducted to evaluate which algorithm is the best accurate at predicting the types of cyber-attacks. DOS Attack, R2L Attack, U2R Attack, and Probe Attack are the four types of attacks we classify. The findings reveal that the suggested machine learning algorithm technique has the best accuracy with entropy calculation, precision, recall, F1 Score, sensitivity, specificity, and entropy calculation.*

Keywords: Cyber-attack, DOS Attack, R2L Attack, U2R Attack, Probe Attack

I. INTRODUCTION

1.1 Data science

This is an interdisciplinary field that employs medical procedures, processes, algorithms, and structures to extract knowledge and insights from unstructured and structured information, as well as to track knowledge and actionable insights across a wide range of utility areas. Peter Naur, who proposed it as a call to action for laptop technology. The International Federation of Classification Societies (IFCS) has been the principal convention to focus on records technology since 1996. Nonetheless, the definition became a work in progress. The term “data era” have become first coined in 2008 thru manner of way of D.J. Patil, and Jeff Hammerbacher, the pioneer leads of data and analytics efforts at LinkedIn and Facebook. In a great deal much less than a decade, it has emerge as one of the most up to date and most trending professions with inside the market. This topic (Data technology I s the sector of study that combines region understanding, programming skills, and understanding ofmathematics and data to extract considerable insights from data. This can be defined as a combination of mathematics, business enterprise acumen, tools, algorithms and tool analyzing techniques, all of which help us in finding out the hidden insights or patterns from raw data which can be of fundamental use with inside the formation of huge business enterprise decisions.

1.2 Data Scientist

This will look at which questions need to be answered and where to find the answers. They have business sense and analytical skills, as well as the ability to extract, clean, and present data. Fact scientists help businesses find, organise, and analyse large amounts of unstructured data.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, May 2022
Impact Factor: 6.252

A. Required skills for a data scientist

- Python, SQL, Scala, Java, R, MATLAB are the programming required.
- Natural Language Processing, Classification, Clustering are Machine Learning techniques required.
- Tableau, SAS, D3.js, Python, Java, R libraries are the Data Visualization techniques required.
- MongoDB, Oracle, Microsoft Azure, Cloudera are the Big data platforms required.

1.3 Artificial Intelligence

This (AI) refers to the simulation of human intelligence in computers that are programmed to act and move like people. The time period can be used to any device that is well-known for displaying improvements in human intellect, such as mastery and problem-solving. Artificial intelligence (AI) refers to intelligence demonstrated by machines rather than human or animal intelligence. Leading AI textbooks define the field as the study of "sensible agents," or any device that is capable of perceiving its environment and acting in ways that increase its chances of achieving its goals. Some well-known debtors use the term "synthetic intelligence" to describe machines that mimic "cognitive" features such as "mastering" and "hassle" that people associate with human thoughts, but Important AI researchers disagree with this definition. Artificial intelligence (AI) is the simulation of human intelligence strategies by computers, especially computer structures. Professional structures, herbal language processing, speech popularity, and device vision are examples of AI packages. Superior internet search engines, advising systems (as used by Youtube, Amazon, and Netflix), understanding human speech (as used by Siri or Alexa), self-driving automobiles (e.g. Tesla), and competing at the highest level in strategic recreation systems are all examples of AI applications (consisting of chess and Go), The AI effect is a phenomena that occurs when computers get more capable. As machines become more capable, obligations that need "intelligence" are typically removed from the definition of AI. For example, optical guy or woman popularity is sometimes overlooked. Fantastically mathematical statistical device mastering reigned the industry in the first many years of the twenty-first century, and this method has proven fantastically successful, assisting in the resolution of many difficult challenges in business and academia. The various subfields of AI research are centred on precise desires and the application of precise tools. Reasoning, knowledge representation, planning, mastery, herbal language processing, belief, and the ability to carry and handle devices are all common AI goals. Some of the area's long-term goals include general intelligence (the ability to solve any problem). To address these difficulties, AI researchers employ strategies such as seek and mathematical optimization, formal logic, synthetic neural networks, and statistics, chance, and economics-based techniques. AI also draws on computer science, psychology, linguistics, philosophy, and a variety of other disciplines. The field was renamed after the assumption that human intelligence "might be so precisely defined that a device could be built to imitate it." This heightens philosophical debates on the ethics and thoughts of creating synthetic beings with human-like intellect. Given the antiquity, these issues were examined through myth, fiction, and philosophy. Science fiction and futurology have also suggested that, due to its enormous capacity and strength, AI could become an existential threat to humans. Carriers were scrambling to sell how their services and products incorporate AI as the hoopla around AI grew. Frequently, what they refer to as AI is certainly one aspect of AI, namely device mastery. For designing and training device mastery algorithms. AI necessitates a foundation of specialised hardware and software. Although no single programming language is synonymous with AI, a handful stand out, including Python, R, and Java. In general, AI systems work by collecting large amounts of classified educational data, analysing the data for correlations and patterns, and then applying those styles to produce predictions about future states. In this method, a chatbot fed instances of textual content chats can study how to produce lifestyles such as face-to-face interactions, or a photo popularity device can learn how to find and define gadgets in images by looking at hundreds of thousands of examples. Mastering, reasoning, and self-correction are three cognitive capabilities that AI programming focuses on. Techniques for learning This branch of AI programming focuses on gathering facts and formulating rules for converting those facts into actionable data. The rules, also known as algorithms, provide computer devices with step-by-step instructions on how to do a specific task. Reasoning techniques. This section of AI programming focuses on choosing the best collection of rules to achieve a desired result. Self-correction techniques.

1.4 Machine Learning

This topic (Machine learning) entails anticipating the future from beyond the records. Machine learning (ML) is a type of artificial intelligence (AI) that allows computer systems to learn without having to be explicitly programmed. Machine

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, May 2022
Impact Factor: 6.252

learning focuses on the creation of computer programmes that may change when exposed to new data, as well as the principles of Machine Learning, such as the design of a simple system learning set of rules using Python. The usage of specialised algorithms is part of the education and prediction process. It feeds the education records to a set of rules, and the set of rules uses the education records to make predictions on new examination records. Machine learning can be classified into a few different categories. There are three types of learning: supervised learning, unsupervised learning, and reinforcement learning. The input records and the accompanying labelling are individually submitted to supervised learning software, and the study records must be categorised by a person beforehand. There are no classifications for unsupervised learning. It provided a set of guidelines for learning. The clustering of the enter records must be parented out by this set of criteria. Finally, Reinforcement Learning dynamically interacts with its surroundings and receives positive or negative feedback to improve its performance. To discover styles in python that result in useful insights, data scientists use a variety of different forms of system learning algorithms. At a high level, those unique algorithms can be divided into two groups based on how they "study" records to make predictions: supervised and unsupervised learning. Classification is a method of estimating the elegance of a set of data items. Occasionally, classes are referred to as goals, labels, or categories. The project of approximating a mapping feature from enter variables(X) to discrete output variables is known as classification predictive modelling (y). Class is a supervised learning technique in system learning and statistics in which the computer software learns from the records input given to it and then applies that learning to categorise fresh observations. This data set can be bi-elegance (for example, determining whether the individual is male or female or whether the mail is unsolicited or non-unsolicited) or multi-elegance (for example, determining whether the individual is male or female or whether the mail is unsolicited or non-unsolicited). Speech recognition, handwriting recognition, biometric identity, report class, and other issues are instances of class issues. As seen in Figure 1.1, this is how it works.

Figure 1.1: Process of Machine Learning

Supervised Machine Learning is used in practically all realistic device research. Supervised researching is when you have input variables (X) and output variables (y), and you utilise a set of rules to investigate the mapping feature from the input to the output, which is y = f. (X). The goal is to approximate the mapping feature to the point where if you have new input data (X), you can anticipate the output variables (y) for that data. Logistic regression, multi-elegance type, Decision Trees, and assist vector machines are examples of supervised machine learning techniques.

II. EXISTING SYSTEM

They proposed first to create a contrastive self-supervised studying to the paradox detection hassle of attributed networks. CoLa, is particularly includes 3 components: contrastive example pair sampling, GNN-primarily based totally contrastive studying version, and multiround sampling-primarily based totally anomaly rating computation.Their version captures the connection among every node and its neighbouring shape and makes use of an anomaly-associated goal to teach the contrastive studying version. We agree with that the proposed framework opens a brand new possibility to extend self-supervised studying and contrastive studying to an increasing number of graph anomaly detection applications. The multiround expected ratings via way of means of the contrastive studying version are in addition used to assess the abnormality of every node with statistical estimation. The schooling segment and the inference segment. In the schooling segment, the contrastive studying version is educated with sampled example pairs in an unmonitored fashion. After that the paradox rating for every node is acquired withinside the inference segment.

III. PROPOSED SYSTEM

The proposed model is to construct a machine learning model for anomaly detection. Anomaly detection is an critical method for spotting fraud activities, suspicious activities, community intrusion, and different peculiar occasions that could have wonderful importance however are hard to detect. The gadget studying version is constructed through making use of right facts technological know-how strategies like variable identity this is the based and unbiased variables. Then the visualisation of the facts is executed to insights of the facts. The model is construct primarily based totally at the preceding

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, May 2022
Impact Factor: 6.252

dataset wherein the set of rules research facts and get skilled one of a kind algorithms are used for highest comparisons. The overall performance metrics are calculated and in comparison.

3.1 Classification of Attacks

The KDD Cup99 data collection contains normal and 22 attack type data with 41 features, and all created traffic patterns are labelled as "normal" or "attack" for further study. There are several types of attacks that penetrate the network over time, and the attacks are divided into four categories.

- Denial of Service (DoS)
- User to Root (U2R)
- Remote to User (R2L)
- Probing

A. Denial of Service

Denial of Service (DoS) is a class of attacks in which an attacker renders a computer or memory resource too busy or too full to handle valid requests, denying legitimate consumers access to a device. The one-of-a-kind approaches for launching a DoS attack are misusing the computer's legitimate functions, concentrating on implementation defects, and exploiting system misconfiguration. DoS attacks are classified based on the services that an attacker makes unavailable to legitimate users.

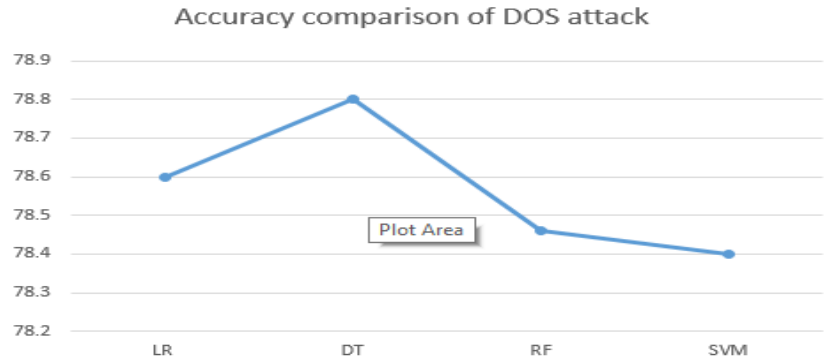

Figure 2: Accuracy obtained in DOS attack

B. User to Root

In a User to Root attack, an attacker gains access to the machine through a regular user account and then gains root access. Regular programming flaws and assumptions about the environment give an attacker the opportunity to exploit the vulnerability of root access.

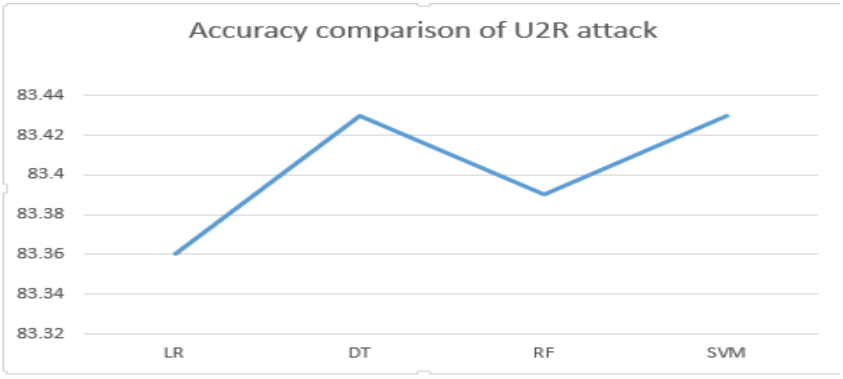

Figure 3: Accuracy obtained in U2R attack

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, May 2022
Impact Factor: 6.252

C. Remote to User

In a Remote to User attack, an attacker sends packets to a device across a network that exploits the machine's vulnerability to gain unauthorised access as a customer. There are various types of R2L attacks, and the most unusual attack on this magnitude is carried out through the use of social engineering.

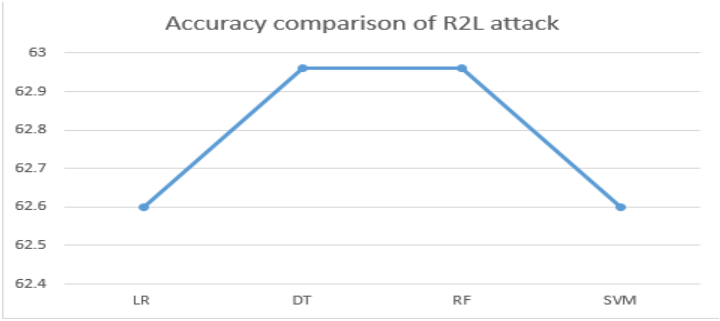

Figure 4: Accuracy obtained in R2L attack

D. Probing

Probing is a type of attack in which an attacker explores a network for information that can be used to find known vulnerabilities. An attacker who has a map of computers and offerings available in a community can manipulate the facts to look for exploits. There are numerous types of probes: some abuse the computer's valid functions, while others employ social engineering techniques. This magnitude of attacks is the most common because it requires little or no technological expertise

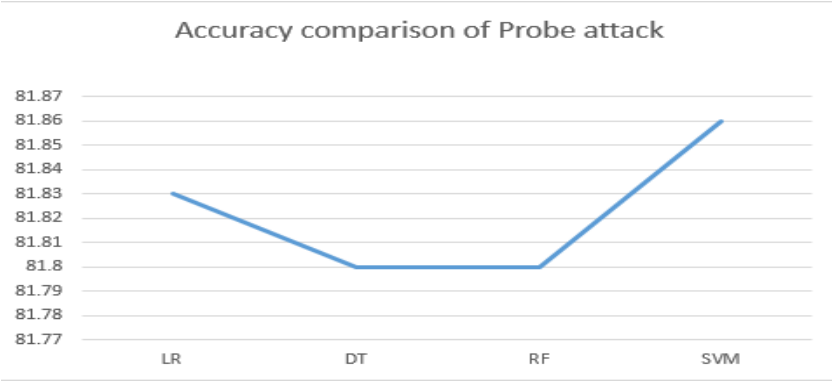

Figure 5: Accuracy obtained in Probe attack

IV. Research And Discussion

The analytical process began with data cleansing and processing, followed by exploratory evaluation, and finally version creation and evaluation. The higher the quality accuracy on the public examination set, the better the accuracy rating. This can be discovered by comparing each set of rules with the type of all community attacks for future prediction outcomes by discovering quality connections. This leads to a number of new insights into diagnosing the community assault of each new connection. To provide a prediction version that uses synthetic intelligence to improve over human accuracy and provide early detection capabilities.

V. Future Work

- The network area must automate the detection of packet transfer attacks from an eligibility perspective (in real time) based on relationship detail.
- To automate this process by displaying the prediction through web software or computer device software.
- To improve the paintings such that they can be used in an Artificial Intelligence context.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, May 2022
Impact Factor: 6.252

REFERENCES

[1]. Wentao Zhao, Jianping Yin and Jun Long , 2008, A Prediction Model of DoS Attack's Distribution Discrete Probability.

[2]. Preetish Ranjan, Abhishek Vaish, 2014, Apriori Viterbi Model for Prior Detection of Socio-Technical Attacks in a Social Network.

[3]. Seraj Fayyad, Cristoph Meinel, 2013, New Attack Scenario Prediction Methodology

[4]. Jinyu W1, Lihua Yin and Yunchuan Guo, 2012, Cyber Attacks Prediction Model Based on Bayesian Network.

[5]. Xiaoyong Yuan , Pan He, Qile Zhu, and Xiaolin Li, 2019, Adversarial Examples: Attacks and Defenses for Deep Learning.

[6]. Wenying Xu, Guoqiang Hu, 2019, Distributed Secure Cooperative Control Under Denial-of-Service Attacks From Multiple Adversaries.

[7]. Zhen Yang, Yaochu Jin, Fellow, and Kuangrong Hao , 2018, A Bio-Inspired Self-learning Coevolutionary Dynamic Multiobjective

[8]. K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?" in Proc. Int. Conf. Learn. Represent., 2019, pp. 1–17.

[9]. Z. Ying, J. You, C. Morris, X. Ren, W. Hamilton, and J. Leskovec, "Hierarchical graph representation learning with differentiable pooling," in Proc. Adv. Neural Inf. Process. Syst., 2018, pp. 4800–4810.

[10]. M. Zhang and Y. Chen, "Weisfeiler-lehman neural machine for link prediction," in Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, Aug. 2017, pp. 575–583.

[11]. T. N. Kipf and M. Welling, "Variational graph auto-encoders," 2016, arXiv:1611.07308.

[12]. G. Pang, C. Shen, L. Cao, and A. van den Hengel, "Deep learning for anomaly detection: A review," 2020, arXiv:2007.02500.

[13]. K. Ding, J. Li, R. Bhanushali, and H. Liu, "Deep anomaly detection on attributed networks," in Proc. SIAM Int. Conf. Data Mining. Philadelphia, PA, USA: SIAM, 2019, pp. 594–602.

[14]. Y. Chen, X. Sean Zhou, and T. S. Huang, "One-class SVM for learning in image retrieval," in Proc. Int. Conf. Image Process., vol. 1, 2001, pp. 34–37.

[15]. X. Xu, N. Yuruk, Z. Feng, and T. A. J. Schweiger, "SCAN: A structural clustering algorithm for networks," in Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD), 2007, pp. 824–833.

[16]. B. Perozzi and L. Akoglu, "Scalable anomaly ranking of attributed neighborhoods," in Proc. SIAM Int. Conf. Data Mining, Jun. 2016, pp. 207–215.

[17]. J. Li, H. Dani, X. Hu, and H. Liu, "Radar: Residual analysis for anomaly detection in attributed networks," in Proc. 26th Int. Joint Conf. Artif. Intell., Aug. 2017, pp. 2152–2158.

[18]. Z. Peng, M. Luo, J. Li, H. Liu, and Q. Zheng, "ANOMALOUS: A joint modeling approach for anomaly detection on attributed networks," in Proc. 27th Int. Joint Conf. Artif. Intell., Jul. 2018, pp. 3513–3519.

[19]. G. Pang, C. Shen, H. Jin, and A. van den Hengel, "Deep weaklysupervised anomaly detection," 2019, arXiv:1910.13601.

[20]. G. Pang, C. Shen, and A. van den Hengel, "Deep anomaly detection with deviation networks," in Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, Jul. 2019, pp. 353–362.

[21]. L. Ruff et al., "Deep one-class classification," in Proc. Int. Conf. Mach. Learn., 2018, pp. 4393–4402.

[22]. Y. Li, X. Huang, J. Li, M. Du, and N. Zou, "SpecAE: Spectral AutoEncoder for anomaly detection in attributed networks," in Proc. 28th ACM Int. Conf. Inf. Knowl. Manage., Nov. 2019, pp. 2233–2236.