# DATA ANALYTICS APPROACH TO CYBERCRIMES USING MACHINE LERNING

**[1]P. Bharath Kumar, [2]N. Eswar reddy, [3]P. Sevarsitha, [4]CH. Aditya Koushik**
**[5]Dr.K. P. Kaliyamurthie**

[1,2,3,4]Students, [5]Guide
Department of Computer Science and Engineering
Bharath Institute of Higher Education and Research-BIHER

***Abstract*- The fast propagation of computer networks has changed the viewpoint of network security. An easy accessibility conditions cause computer network as susceptible against several threats from hackers. Threats to networks are numerous and potentially devastating. Up to the moment, researchers have developed Intrusion Detection Systems (IDS) capable of detecting attacks in several available environments. A boundlessness of methods for misuse detection as well as anomaly detection has been applied. Many of the technologies proposed are complementary to each other, since for different kind of environments some approaches perform better than others. This project presents a new intrusion detection system that is then used to survey and classify them. The taxonomy consists of the detection principle, and second of certain operational aspects of the intrusion detection**

***Keywords*: Machine Learning, Data Set.**

**OBJECTIVE**

The goal of intrusion detection is to monitor the network assets to detect anomalous behavior and misuse in network. Intrusion detection concept was introduced in early 1980's after the evolution of internet with surveillance end monitoring the threat. There was a sudden rise in reputation and incorporation in security infrastructure. Since then, James Anderson's wrote a paper for a government organization and imported an approach that audit trails contained important information that could be valuable in tracking misuse and understanding of user behavior. Then the detection appeared and audit data and its importance led to terrific improvements in the subsystems of every operating system. IDS and Host Based Intrusion Detection System (HIDS) were first defined. In 1983, SRI International and Dorothy Denning began working on a government project that launched a new effort into intrusion detection system development. Around 1990s the revenues are generated and intrusion detection market has been raised. Real secure is an intrusion detection network developed by ISS. After a year, Cisco recognized the priority for network intrusion detection and purchased the Wheel Group for attaining the security solutions. The government actions like Federal Intrusion Detection Networks (FID Net) were designed under Presidential Decision Directive 63 is also adding impulse to the IDS .

The main objectives of the problem are as follows

1. Less throughput.
2. No energy checkup after each communication.
3. Time delay is high.
4. High packet loss.

## I.INTRODUCTION

In the statistical context, Machine Learning is defined as an application of artificial intelligence where available information is used through algorithms to process or assist the processing of statistical data. While Machine Learning involves concepts of automation, it requires human guidance. Machine Learning involves a high level of generalization in order to get a system that performs well on yet unseen data instances.

Machine learning is a relatively new discipline within Computer Science that provides a collection of data analysis techniques. Some of these techniques are based on well established statistical methods (e.g. logistic regression and principal component analysis) while many others are not.

Most statistical techniques follow the paradigm of determining a particular probabilistic model that best describes observed data among a class of related models. Similarly, most machine learning techniques are designed to find models that best fit data (i.e. they solve certain optimization problems), except that these machine learning models are no longer restricted to probabilistic ones. Therefore, an advantage of machine learning techniques over statistical ones is that the latter require underlying probabilistic models while the former do not. Even though some machine learning techniques use probabilistic models, the classical statistical techniques are most often too stringent for the oncoming Big Data era, because data sources are increasingly complex and multi-faceted. Prescribing probabilistic models relating variables from disparate data sources that are plausible and amenable to statistical analysis might be extremely difficult if not impossible.

Machine learning might be able to provide a broader class of more flexible alternative analysis methods better suited to modern sources of data. It is imperative for statistical agencies to explore the possible use of machine learning techniques to determine whether their future needs might be better met with such techniques than with traditional ones.

## II.LITERATURE REVIEW

**NL-IDS: Trust Based Intrusion Detection System for Network layer in Wireless Sensor Networks Umashankar Ghugar, Jayaram Pradhan IEEE 2021.**

From the last few years, security in wireless sensor network (WSN) is essential because WSN application uses important information sharing between the nodes. There are large number of issues raised related to security due to open deployment of network. The attackers disturb the security system by attacking the different protocol layers in WSN. The standard AODV routing protocol faces security issues when the route discovery process takes place. The data should be transmitted in a secure path to the destination. Therefore, to support the process we have proposed a trust based intrusion detection system (NL-IDS) for network layer in WSN to detect the Black hole attackers in the network. The sensor node trust is calculated as per the deviation of key factor at the network layer based on the Black holeattack. in the network. The sensor node trust is calculated as per the deviation of key factor at the network layer based on the Black hole attack. We use the watchdog technique where a sensor node continuously monitors the neighbor node by calculating a periodic trust value. Finally, the overall trust value of the sensor node is evaluated by the gathered values of trust metrics of the network layer (past and previous trust values). This NL-IDS scheme is efficient to identify the malicious node with respect to Black hole attack at the network layer. To analyze the performance of NL-IDS, we have simulated the model in MATLAB R2015a, and the result shows that NL-IDS is better than Wang et al. [11] as compare of detection accuracy and false alarm rate.

**Analyzing the Vulnerability of Wireless Sensor Networks to a Malicious Matched Protocol Attack George D. O'Mahon, Philip J. Harris, Colin C. Murphy IEEE 2021.**
Safety critical, Internet of Things (IoT) and space-based applications have recently begun to adopt wireless networks based on commercial off the shelf (COTS) devices and standardized protocols, which inherently establishes the security challenge of malicious intrusions. Malicious intrusions can cause severe consequences if undetected, including, complete denial of services. Particularly, any safety critical application requires all services to operate correctly, as any loss can be detrimental to safety and/or privacy. Therefore, in order for these safety critical services to remain operational and available, any and all intrusions need to be detected and mitigated. Whilst intrusion detection is not a new research area, new vulnerabilities in wireless networks, especially wireless sensor networks (WSNs), can be identified. In this paper, a specific vulnerability of WSNs is explored, termed here the matched protocol attack. This malicious attack uses protocol-specific structures to compromise a network using that protocol. Through attack exploration, this paper provides evidence that traditional spectral techniques are not sufficient to detect an intrusion using this style of attack. Furthermore, a ZigBee cluster head network, which co-exists with ISM band services, consisting of XBee COTS devices is utilized, along with a real time spectrum analyzer, to experimentally evaluate the effect of matched protocol interference on a realistic network model. Results of this evaluation are provided in terms of device errors and spectrum use. This malicious challenge is also examined through Monte-Carlo simulations. A potential detection technique, based on coarse inter-node distance measurements, which can theoretically be used to detect matched protocol interference and localize the origin of the source, is also suggested as a future progression of this work. Insights into how this attack style preys on some of the main security risks of any WSN (interoperability, device

## III. METHODOLOGY

The IDS have been implemented in organizations to collect and analyze various types of attacks within a host system or a network. In addition, to identify and detect possible threats violations, which involve both intrusions, which are the attacks from outside the organizations and misuses that are known as the attacks within the organizations. In this paper, we proposed the integrated model which involves a combination of the two systems Intrusion Detection (ID) and Intrusion Prevention (IP) adding to those getting benefits from well-known techniques: intruder Detection (ID) which is totally different from most of the recent works that focused only on using one system, either detection or prevention and also using either Intruder detection or Signature based detection. Some works even used a hybrid method which is a combination of both such as the work presented where the researchers used ID based on Signature but even then, their method was not provided with prevention capabilities. On the other hand, in our case, we proposed to use our approach IDPS, which not only can detect the attack

## IV. Existing System

**Naïve Bayes (NB):** In statistics, **naive Bayes classifiers** are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naïve) independence assumptions between the features. They are among the simplest Bayesian network models, but coupled with kernel density estimation, they can achieve higher accuracy levels.
Naïve Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression, which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers.
In the statistics and computer science literature, naive Bayes models are known under a variety of names, including **simple Bayes** and **independence Bayes**. All these names reference the use of Bayes' theorem in the classifier's decision rule, but naïve Bayes is not (necessarily) a Bayesian method.
Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. There is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable.

## V. Proposed System

**Decision Tree (DT):** A **decision tree** is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

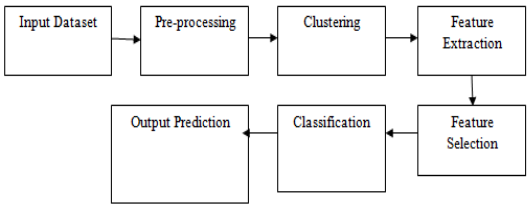
Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal, but are also a popular tool in machine learning.

A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules.

In decision analysis, a decision tree and the closely related influence diagram are used as a visual and analytical decision support tool, where the expected values (or expected utility) of competing alternatives are calculated.

A decision tree consists of three types of nodes

1. Decision nodes – typically represented by squares
2. Chance nodes – typically represented by circles
3. End nodes – typically represented by triange




**SYSTEMARCHITECTURE**


**Use Case Diagram**

**SYSTEM IMPLEMENTATION**

**MODULE DESCRIPTION**

There are 8 components in the system. They are

- Incoming packet
- Packet Capture
- Packet scanner
- Packet analyzer
- Labeling
- Training model
- Prediction:
- Output

**COMPARATIVE STUDY OF EXISTING AND PROPOSED SYSTEM**

In our project we have used algorithms like Naïve Bayes (NB) and Decision Tree (DT). All are measured in terms of accuracy. From the result it is proved that the proposed Decision Tree (DT) works better than other existing algorithm. All are measured in terms of accuracy. From this we are getting good accuracy so we can state that our proposed system works better than the existing system.

**Future Scope**

Future research should consider other machine learning algorithms to ascertain more efficient ways to perform the classification technique on the datasets. It is recommended that further research should be carry out on other parameters that can improve the accuracy of detection

**ALGORITHM OF PROPOSED WORK**

i. Dataset is divided into small overlapping or either non-overlapping blocks.
ii. Extract the features using traditional techniques.
iii. Extracted feature values corresponding to each key point are stored in matrix.
iv. Apply sorting techniques to get similar features that lie in nearness.
v. Introduce shift vector concept to find key point with similar shifting.
vi. Use the counter vector to count the occurrence same shifting key point and set the counter to
vii. Similar regions are identified with the help of threshold value above steps are used.

**HARDWAREREQUIREMENTS**

- Processor: Core I5 Processor.
- Ram: 4 GB RAM
- Hard Disk: 500 G.B Hard Disk
- 14 inch monitor

**SOFTWAREREQUIRMENTS**

- Technology: Python
- IDE : Python IDE
- WebServer: Jupyter/Anaconda/Panda
- Database : My SQL

**REFERENCES:**

[1] I. F. Akyildiz et al., “Wireless Sensor Networks A Survey, “Elsevier Comp. Networks, vol. 3, no. 2, 2019, pp. 393–422
[2] G.Li, J.He, Y. Fu. “Group-based intrusion detection system in wireless sensor networks” Computer Communications, Volume 31, Issue 18 (December 2019)
[3] Michael Brownfield, “Wireless Sensor Network Denial of Sleep Attack”, Proceedings of the 2019 IEEE Workshop on Information Assurance and Security United States Military Academy, West Point, NY.
[4] FarooqAnjum, DhanantSubhadrabandhu, SaswatiSarkar *, Rahul Shetty, “On Optimal Placement of Intrusion Detection Modules in Sensor Networks”, Proceedings of the First International Conference on Broadband Networks (BROADNETS19).
[5] Parveen Sadotra et al, International Journal of Computer Science and Mobile Computing, Vol.5 Issue.9, September- 2019, pg. 23-28
[6] K. Akkayaand M. Younis, ―A Survey of Routing Protocols in Wireless Sensor Networks, ‖ in the Elsevier Ad Hoc Network Journal, Vol. 3/3 pp. 325-349, 2019.
[7] A. Abduvaliyev, S. Lee, Y.K Lee, “Energy Efficient Hybrid Intrusion Detection System for Wireless Sensor Networks”, IEEE International Conference on Electronics and Information Engineering, Vol.2, pp. 25-29, August 2019.
[8] Parveen Sadotra and Chandrakant Sharma. A Survey: Intelligent Intrusion Detection System in Computer Security. International Journal of Computer Applications 151(3):18-22, October 2019.
[9] A. Araujo, J. Blesa, E. Romero, D. Villanueva, “Security in cognitive wireless sensor networks. Challenges and open problems”, EURASIP Journal on Wireless Communications and Networking, February 2019.
[10] A. Becher, Z. Benenson, and M. Dorsey, \Tampering with motes: Real-world physical attacks on wireless sensor networks." in SPC (J. A. Clark, R. F. Paige, F. Polack, and P. J.Brooke, eds.), vol. 3934 of Lecture Notes in Computer Science, pp. 104{118, Springer, 2019.
[11] I. Krontiris and T. Dimitriou, \A practical authentication scheme for in-network programming in wireless sensor networks," in ACM Workshop on Real- World Wireless Sensor Networks, 2019.
[12] M. Ali Aydın *, A. HalimZaim, K. GokhanCeylan “A hybrid intrusion detection system design for computer network security” Computers and Electrical Engineering 35 (2019)