A Study of Network Intrusion Detection using Machine Learning

Ms. Reena Ostwal[1*]; Dr. Anil Pimpalapure[2]

Eklavya University Damoh, University in Padriya, Madhya Pradesh
*Corresponding Author

Received: 05 June 2024/ Revised: 12 June 2024/ Accepted: 20 June 2024/ Published: 30-06-2024

Abstract— *Network security engineers work to keep services available all the time by handling intruder attacks. Intrusion Detection System (IDS) is one of the obtainable mechanisms that is used to sense and classify any abnormal actions. Therefore, the IDS must be always up to date with the latest intruder attacks signatures to preserve confidentiality, integrity, and availability of the services. The speed of the IDS is a very important issue as well learning the new attacks. This research work illustrates how the Knowledge Discovery and Data Mining (or Knowledge Discovery in Databases) KDD dataset is very handy for testing and evaluating different Machine Learning Techniques. It mainly focuses on the KDD preprocess part in order to prepare a decent and fair experimental data set. The J48, MLP, and Bayes Network classifiers have been chosen for this study. It has been proven that the J48 classifier has achieved the highest accuracy rate for detecting and classifying all KDD dataset attacks, which are of type DOS, R2L, U2R, and PROBE.*

Keywords— Network security, Intrusion Detection System, Knowledge Discovery, Databases.

I. INTRODUCTION

BUILDING a reliable network is a very difficult task considering all different possible types of attacks. Nowadays, computer networks and their services are widely used in industry, business, and all arenas of life. Security personnel and everyone who has a responsibility for providing protection for a network and its users, have serious concerns about intruder attacks. Network administrators and security officers try to provide a protected environment for users' accounts, network resources, personal files and passwords. Attackers may behave in two ways to carry out their attacks on networks; one of these ways is to make a network service unavailable for users or violating personal information. Denial of service (DoS) is one of the most frequent cases representing attacks on network resources and making network services unavailable for their users. There are many types of DoS attacks, and every type has it is own behavior on consuming network resources to achieve the intruder's aim, which is to render the network unavailable for its users [1]. Remote to user (R2L) is one type of computer network attacks, in which an intruder sends set of packets to another computer or server over a network where he/she does not have permission to access as a local user. User to root attacks (U2R) is a second type of attack where the intruder tries to access the network resources as a normal user, and after several attempts, the intruder becomes as a full access user [2]. Probing is a third type of attack in which the intruder scans network devices to determine weakness in topology design or some opened ports and then use them in the future for illegal access to personal information. There are many examples that represent probing over a network, such as map, port sweep, ips-weep. IDS become an essential part for building computer network to capture these kinds of attacks in early stages, because IDS works against all intruder attacks. IDS uses classification techniques to make decision about every packet pass through the network whether it is a normal packet or an attack (i.e. DOS, U2R, R2L, and PROBE) packet. KDD is an online repository dataset, which includes all types of intruders' attacks such as DOS, R2L, U2R, and PROBE. In this paper, a number of classifiers will be evaluated on the KDD dataset. The methodology followed in this study is first to perform a preprocessing step on KDD dataset and after to use the prepared dataset on a fair environment and resources, and finally, to examine which classifier is more accurate than others in detecting all studied attacks (DOS, R2L, U2R, and PROBE).

1.1 Classification in Data Mining:

In Data mining, order is planned to make a figure of the participations in a gathering for Data occurrences. This procedure uses complex investigation of Data to decide Data organizations in immense datasets. Because of its perplexing highlights, therapeutic databases give complexities to design blackmail.

There are two ways to deal with Data mining: measurable and AI algorithms. The procedures in Data mining are characterized into engaging and prescient. Spellbinding mining assignments give the general Data properties in the database. For Predictive mining assignments, derivation is made on the Data for predictions [9] whereby conjecture is made on express qualities dependent on examples recognized by known outcomes. Enlightening Data mining, without having any predefined target, gives attributes and depictions to the Data collection.

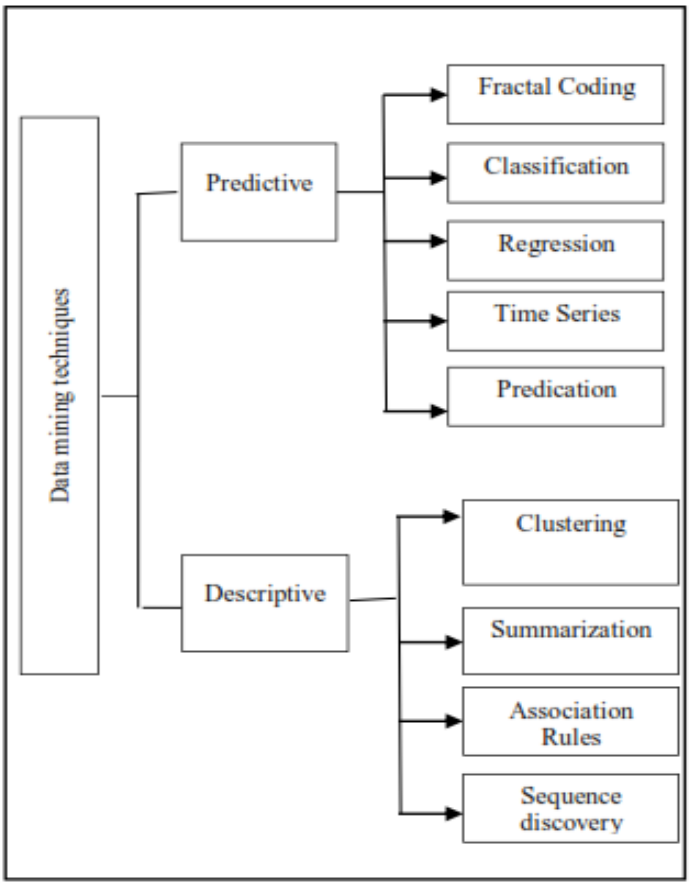

FIGURE 1: Data mining methodology

Data mining systems are viable and prescient for future examples in light of the fact that:

a) It is easy to use and prediction depends on past conditions

b) It works by gaining from past Data

c) Data from various assets is overseen and just required Data is extricated d) models are effectively refreshed by knowledge, past data and change in patterns.

These are what make it dependable and common sense in the medicinal picture order.

The three knowledge approaches in Data mining algorithms are managed (the calculation works with a lot of models with known marks whose qualities are ostensible in arrangement task, or numerical in relapse task), unsupervised (obscure names in the dataset, and the calculation commonly goes for gathering instances as per the closeness of their trait esteems, portraying a classification undertaking, or semi-regulated (whereby knowledge is led when there is accessibility of a little subset of named instances, simultaneous with a substantial number of unlabelled examples).The errand of sorting is viewed as a directed strategy in which each case has a place with a class, indicated by the estimation of a unique objective property or the class characteristics.

II. COMPARATIVE ANALYSIS OF DATA MINING METHODOLOGIES:

Arrangements came about because of classification calculation are excellent yet starting at now, none is different and adaptable to be acknowledged for the most part in the therapeutic Data order network. Straight out factors in medicinal Data are every so often valuable to land at decisions and to sum up data. All out Data (for example characterization of infection and non-illness gatherings) is helpful for Data mining method and furthermore simple to separate restorative data.

Data mining strategies, which are an on-going application in the restorative space, are connected in mining therapeutic Data, which involves association rule digging for finding regular examples, prediction, characterization and classification. To date, there have been many researches on this and know-how and decision emotionally supportive networks have been created to make increasingly exact analysis and prediction of infections particularly in anticipating heart ailments, lung and bosom malignant growth and remote wellbeing checking.

The act of utilizing solid Data and proof to help restorative decisions (otherwise called proof based prescription or EBM) has existed for quite a long time. John Snow, viewed as the dad of present day the study of disease transmission, utilized maps with early types of visual diagrams in 1854 to find the wellspring of cholera and demonstrate that it was transmitted through the water supply, underneath (Tufte 1997, Audain 2007).

2.1 Network Intrusion Detection Systems:

System interruption locations frameworks have turned into a standard segment in PC arrange security backgrounds as they permit organize directors to distinguish approach infringement. These arrangement infringement may go from outside aggressors endeavouring to increase unapproved get to (which can for the most part be ensured against through the remainder of the security background) to insiders manprocessing their entrance (which as a rule isn't anything but difficult to ensure against). Recognizing such infringement is an essential advance in making remedial move.

Then again, recognizing approach infringement enables executives to distinguish zones where their barriers need improvement, for example, by recognizing a formerly obscure powerlessness, “a framework that wasn't appropriately fixed, or a client that needs further instruction against social designing assaults”

“The issue is that present NIDS are tuned explicitly to recognize realized administration level system assaults. Endeavors to extend past this restricted domain commonly results in an unsatisfactory dimension of false positives. In the meantime, enough Data exists or could be gathered to enable executives to recognize these approach infringement”

Shockingly, Data is substantial, examine procedure so tedious, that overseers don't have the assets to experience everything and locate the important knowledge, put something aside for the most uncommon circumstances, for example, after the association has assumed a huge misfortune and the investigation is done as a major aspect of a legitimate examination. At the end of the day, arrange overseers don't have the assets to proactively break down the Data for approach infringement, particularly within the sight of a number of false positives that reason them to squander their constrained assets.

Given the idea of this issue, the common arrangement is Data-mining/AI in a disconnected domain. Such a methodology would add extra profundity to the heads resistances, and enables them to all the more precisely figure out what the dangers against their system are however the utilization of various methodologies on Data from numerous sources.

Consequently, movement that it isn't productive to distinguish in close continuous in an online Network Intrusion Detection (NID), either because of the measure of express that would should be held or the measure of computational assets that would should be used in a restricted time window, can be all the more effectively recognized.

A several instances of what such a framework could recognize, that online Network Intrusion Detection System (NIDS) can't identify adequately, incorporate specific sorts of vindictive movement, for example, low and moderate sweeps, a gradually engendering worm, irregular action of a client dependent on some new example of action (instead of a solitary association or modest number of organizations, which will undoubtedly create various false positives) or indeed, even new sorts of ambushes that online sensors are not tuned for.

Also, such “a framework could all the more effectively take into account the presentation of new measurements that can utilize the verifiable Data as a gauge for examination with current action. It additionally serves to help organize overseers, security officers, and experts in the execution of their obligations by enabling them to make inquiries that would not have jumped out at them from the earlier” In a perfect world, such a framework ought to have the capacity to infer a risk level for the system action that it breaks down, and foresee future assaults dependent on past movement.

In my examination, the focus is on the mining of system association Data as an initial step. System association Data is anything but difficult to gather from firewalls and online-system interruption-sensors, or it very well may be built dependent on parcel

logs. It introduces less legitimate issue than different types of Data that could be gathered in numerous situations since it doesn't distinguish clients (just machines), it doesn't contain subtleties of what was done, and it is effectively-anonym sable.

2.2 Objective of System Interruption Discovery:

So as to make sense of how Data mining/AI can be connected to discover applicable PC security data, we should initially characterize what Data mining and AI are. By and large, Data mining is the way toward removing helpful and beforehand unnoticed models or examples from extensive Data stores. Data mining is a segment of the Knowledge Discovery in Databases (KDD) process.

AI is a logical order that is worried about the plan and advancement of algorithms that enable PCs to advance practices dependent on observational Data, for example, from sensor Data or databases. This work will address some different segments of that procedure, for example, highlight determination, and essentially worried about the Data mining and AI systems.

Data mining strategies can be separated by their diverse model capacities and portrayal, inclination paradigm, and algorithms the principle capacity of the model that we are keen on is classification, as typical, or vindictive, or as a specific kind of assault.

Also, Data mining frameworks give the way to effectively perform Data synopsis and representation, helping the security expert in distinguishing zones of concern. The models must be spoken to in some structure. Basic portrayals for Data mining systems incorporate standards, decision trees, straight and non-direct capacities (counting neural nets), time based instances, and likelihood models.

These portrayals can be utilized when digging for security Data. While a portion of the work in study utilizes different inclination measure, for example, Receiver Operating Characteristic Curve (ROC Curve), and exactness, the principally concerned is with the precision. The model portrayal, digging for security knowledge utilizes various inquiry algorithms, for example, statistical examination, deviation investigation, rule enlistment, neural snatching, making associations, connections, and classification.

Given every one of these methods to discover "shrouded designs dependent on beforehand undetected interruptions help to grow new discovery formats," This enables us to rise above the restrictions of numerous present IDS which depend on a static arrangement of interruption marks (abuse location frameworks) and "advance from remembrance to speculation,". Such a framework would be generally a kind of abnormality discovery framework. "Abnormality discovery endeavors to measure the standard or adequate conduct and banners other sporadic conduct as possibly meddlesome"

The main case of such a framework was IDES, as portrayed in Denning's fundamental paper (1987), which concentrated on mining statistical measures to use for examination while hunting down inconsistencies.

IDS an idea initially presented by Anderson [1] and later formalized by Denning [2] have gotten expanding consideration in the course of recent years. IDSs are frameworks that go for identifying interruptions, i.e., sets of activities that endeavor to bargain the trustworthiness, secrecy or accessibility of a PC asset.

To put it clearly, PC security manages the security of Data and the processing assets and is generally connected with the accompanying three properties:

- **Confidentiality:** It is anticipation of any deliberate or accidental unapproved divulgence of Data. For instance, a gatecrasher finding out about the client Visa database or gaining admittance to the exclusive source code is viewed as a rupture of classification. Note that commonly such a rupture is irreversible and can't be kept effectively.
- The term classification can likewise be comprehended in a more extensive setting in which it additionally relates to the non-conveyance of administrations to unapproved clients, despite the fact that this would not bargain secrecy in itself.
- **Integrity:** It is counteractive action of purposeful or accidental unapproved change of Data. For instance, an interloper damaging the organization's web server or changing the bank's database text for individual increase is an assault against Data uprightness. Note that regularly honesty can be re-established, e.g., from different sources, for example, reinforcement duplicates, in spite of the fact that this procedure might be expensive, tedious, and not constantly total.
- **Availability:** It is a version of the unapproved retaining of registering assets. Instances of accessibility incorporate the refusal of-administration (DoS) assault, in which the assailant hinders the registering assets with the goal that

approved clients can't utilize them, or physical hardware robbery. In light of this meaning of the C.I.A group of three, it tends to be characterized interruption as pursues:

Interruption is any arrangement of activities that endeavor to bargain the secrecy, uprightness or accessibility of a PC asset.

An interruption identification framework screens PC frameworks and systems to decide whether a pernicious time (i.e., an interruption) has happened. Each time a vindictive time is recognized, the IDS raise a caution.

Normally, the prerequisites for classification, respectability and accessibility are not outright, however are characterized by a security strategy.

The security approach states which data is secret, who is approved to adjust given data and what sort of utilization of PC frameworks is satisfactory. In this manner we can reformulate the underlying meaning of interruption as Intrusion is an infringement of a security strategy.

One may classify interruption recognition frameworks as far as conduct i.e., they might be latent (those that essentially create cautions and log organize parcels). They may likewise be dynamic which implies that they recognize and react to assaults, endeavor to fix programming openings before getting hacked or act proactively by logging out potential gate-crashes, or blocking administrations.

2.3 Classification of interruption recognition frameworks:

Fundamentally, AN IDS is worried about the discovery of unfriendly activities. This system security device utilizes both of two primary strategies.

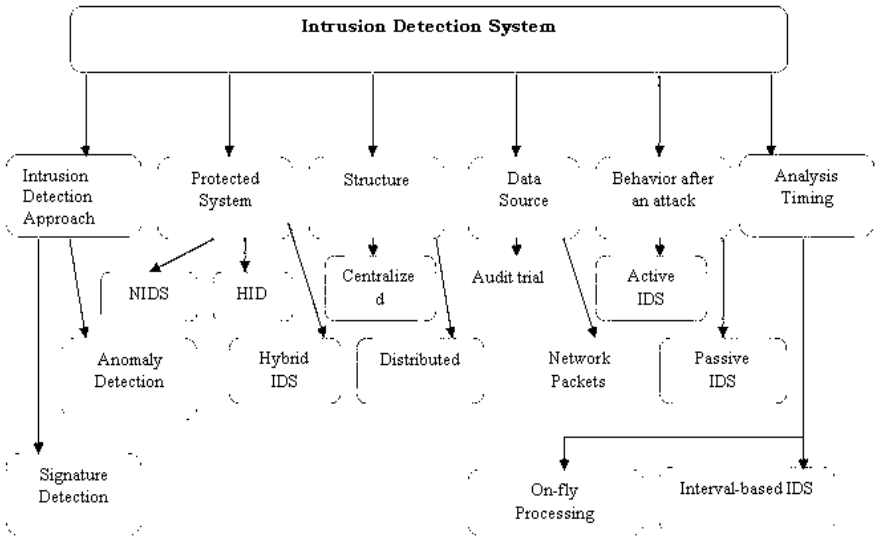

FIGURE 2: Intrusion detection system classification and processing

2.3.1 Intrusion Detection Approach:

This system security instrument utilizes both of two principle procedures (portrayed in more detail beneath). The first, inconsistency identification, investigates issues in interruption location related with deviations from typical framework or client conduct. The second utilizes signature recognition to separate between abnormality or assault designs (marks) and known interruption discovery marks. The two strategies have their unmistakable favourable circumstances and drawbacks just as appropriate application zones of interruption discovery.

2.3.2 Anomaly-Based Detection:

Abnormality based identification is the way toward looking at meanings of what movement is viewed as typical against watched times to distinguish noteworthy deviations. An IDS utilizing peculiarity based recognition has profiles that speak to the ordinary conduct of such things as clients, has, arrange organizations, or applications. The profiles are created by observing the qualities of ordinary action over some undefined time frame. For instance, a profile for a system may demonstrate that Web movement involves a normal of 13% of system transmission capacity at the Internet outskirt amid run of the mill working day hours.

The IDS utilizes measurable strategies to contrast the attributes of current action with limits identified with the profile, for example, recognizing when Web movement involves essentially more data transfer capacity than anticipated and cautioning a head of the irregularity. Profiles can be produced for some conduct properties, for example, the quantity of messages sent by a client, the quantity of fizzled login endeavours for a host, and the dimension of processor utilization for a host in a given timeframe.

2.3.3 Signature-Based Detection:

A mark is an example that relates to a known danger. Mark based recognition is the way toward contrasting marks against watched times with recognize conceivable episodes.

Mark based recognition is powerful at identifying known dangers however generally ineffectual at recognizing already obscure dangers, dangers masked by the utilization of avoidance methods, and numerous variations of known dangers. Mark based discovery is the least difficult location strategy since it just thinks about the present unit of action, for example, a bundle or a log section, to a rundown of marks utilizing string examination activities. Mark based location innovations have small comprehension of many system or application conventions and can't follow and comprehend the condition of complex correspondences.

2.4 Types of Protected Systems:

There are numerous sorts of IDS advances. They are separated into the accompanying three gatherings dependent on the sort of times that they screen and the manners by which they are conveyed:

- **Network Based System:** It screens arrange traffic for specific system sections or gadgets and examines the system and application convention movement to recognize suspicious action.

  It can distinguish a wide range of kinds of times of intrigue. It is most normally sent at a limit between systems, for example, in closeness to fringe firewalls or switches, virtual private system (VPN) servers, remote access servers, and remote systems.

- **Host Based System:** Host-Based System screens the attributes of a solitary host and the times happening inside that have for suspicious action. Instances of the kinds of attributes have based IDS may screen are arrange traffic (just for that have), framework logs, running procedures, application movement, record access and alteration, and framework and application design changes. Host-based IDSs are most ordinarily sent on basic has, for example, freely open servers and servers containing delicate data.

- **Hybrid Based System:** It has been analysed the distinctive IDSs utilize diverse instruments to flag or trigger alerts on your system. It is likewise analysed two areas that IDSs use to look for meddlesome movement. Every one of these methodologies has advantages and disadvantages. By consolidating different methods into a solitary cross breed framework, notwithstanding, it is conceivable to make IDS that has the advantages of various methodologies, while defeating a considerable lot of the disadvantages.

2.5 Structure of IDS:

As for where and how Data is handled by the interruption recognition framework, the interruption location frameworks can be ordered into appropriated and incorporated. A disseminated interruption recognition framework (DIDS) is one where Data is gathered and broke down in numerous hosts, instead of an incorporated interruption location framework (CIDS), in which Data might be gathered in a circulated manner, yet is handled halfway. Both dispersed and concentrated interruption identification frameworks may utilize host-or system based Data accumulation methodologies, or a blend of them.

2.6 Data Source:

Interruption recognition frameworks can keep running on either a consistent or intermittent feed of data (Real-time IDS and Interval-based IDS individually), and consequently, they utilize two diverse interruption discovery approaches. Review trail examination is the common methodology utilized by temporally worked frameworks (or time-based frameworks or intermittently worked frameworks). Interestingly, progressively deployed IDS (or IDS deployed in progressive situations or

IDS intended for continuous deployment) are intended for web-based observing and analyzing (or dissecting) system times and client activities.

2.7 Behaviour of an assailant:

Interruption location frameworks must be equipped for recognizing ordinary (non-security) and anomalous client activities, to find noxious endeavors in time. However, interpreting client practices (or a total client system session) to reach a consistent security-related conclusion is quite complex (or not very simple) — numerous behavioral patterns (or user behavior) are eccentric and vague.

So as to arrange moves, interruption identification frameworks exploit the irregularity discovery approach, now and again alluded to as conduct based [Deb99] or assault marks for example a distinct material on known strange conduct (signature location), additionally called knowledge based.

2.8 Behaviour of the user in the system:

One may classify interruption recognition frameworks as far as conduct i.e., they might be latent (those that just create cautions and log arrange parcels). They may likewise be dynamic which implies that they recognize and react to assaults, endeavor to fix programming openings before getting hacked or act proactively by logging out potential interlopers, or blocking administrations.

2.9 Analysis Timing:

Interruption identification frameworks can keep running on either a persistent or temporal feed of data (Real-time IDS and Interval-based IDS individually) and consequently they utilize two distinctive interruption location approaches. Review trail examination is the common methodology utilized by intermittently worked frameworks. Conversely, the IDS deployable continuously conditions are intended for web based checking and dissecting framework times and client activities.

2.10 Audit Trail Processing:

There are numerous issues identified with review trail (time log) Holbrook A (2003) preparing. Putting away review trail reports in a solitary document must be maintained a strategic distance from since interlopers may utilize this element to roll out undesirable improvements. It is obviously better to keep a specific number of time log duplicates spread over the system; however it would infer adding a several overheads to both the framework and system. Further, from the usefulness perspective, recording all time's imaginable methods a recognizable utilization of framework assets (both the nearby framework and system included). Log pressure, rather, would expand the framework load. Determining which times are to be examined is troublesome on the grounds that specific sorts of assaults may pass undetected. It is additionally hard to foresee how substantial review records can be – through experience one can just make an unpleasant gauge. Additionally, a fitting setting of a capacity period for current review documents is certifiably not a direct assignment. When all is said in done, this relies upon a particular IDS arrangement and its relationship motor. Absolutely, file documents ought to be put away as duplicates for recovery investigation purposes.

2.11 On-Fly Processing:

With on the fly preparing Frank, E. (2005), IDS performs online confirmation of framework times. By and large, a surge of system parcels is always checked continually. With this kind of processing, interruption discovery utilizes the Data of current exercises over the system to detect conceivable assault endeavors (it doesn't search for effective assaults previously). Given the calculation unpredictability, the algorithms that are utilized here are constrained to fast and productive strategies that are frequently algorithmically basic. This is because of a trade-off between the principle essential – assault identification capacity and the multifaceted nature of Data preparing components utilized in the recognition itself. In the meantime, development of an on-the-fly processing IDS device requires a lot of RAM (supports) since no Data stockpiling is utilized. Accordingly, IDS may at some point miss bundles, in light of the fact that practical processing of an excessive number of parcels isn't accessible. The measure of Data gathered by the finder is little since it sees just cushion substance. Consequently, just little parts of data can be investigated for looking through specific qualities or arrangements.

2.12 Intrusion Detection Systems:

We utilized this methodology for answer for characterizing the records as interlopers or typical records. A firewall might be utilized in both home and business situations, Intrusion Detection Systems are just extremely doable inside trade. They are

commonly costly frameworks which are frequently named "Know-how Firewalls". The explanation behind this is they use Data Mining or Machine Knowledge methods to screen designs in system movement. This observing procedure is led to identify deceptive action designs. For instance, on the off chance that we screen the system movement in an organization, in which all clients sign on at 9am and log off at 5pm, at that point if a client signs on at 3am, this is probably going to be an interruption. To accomplish this, Data Mining Models are actualized and prepared on utilization Data. Figure demonstrates the suggested system joining for an interruption discovery framework. The mix of the Intrusion Detection System and a firewall will permit most extreme separating of system traffic and will keep most of assaults. There are some critical focuses that ought to be viewed as when coordinating IDS with a system. The blend of the Intrusion Detection System and a firewall will permit most extreme sifting of system traffic and will keep most of assaults. There are some imperative focuses that ought to be viewed as when coordinating IDS with a system.

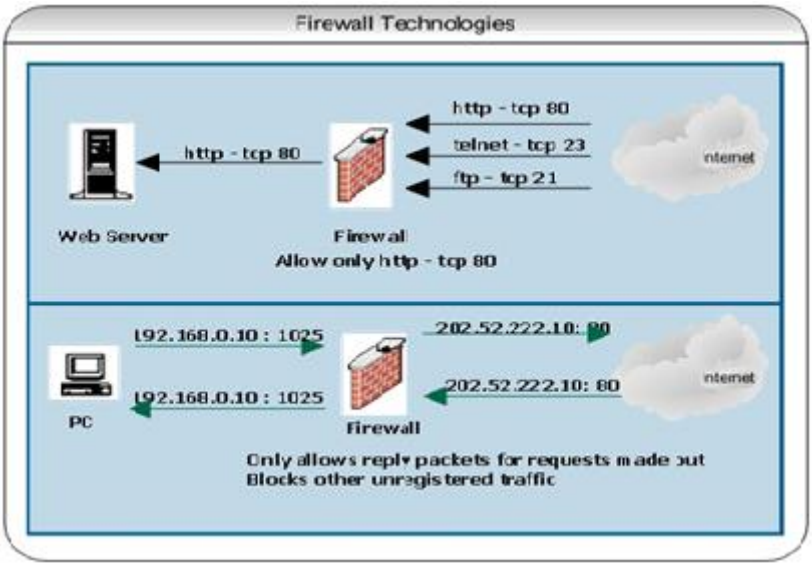

FIGURE 3: Packet filtering (Top) vs. State full Inspection

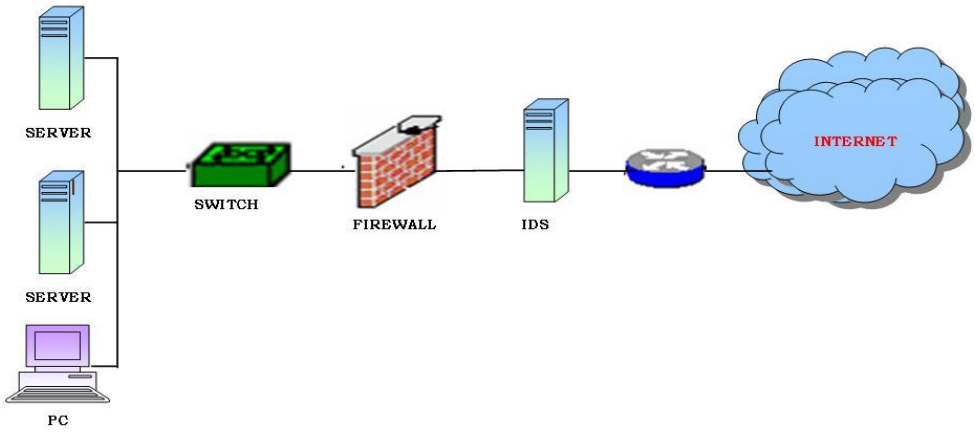

FIGURE 4: Typical Intrusion detection systems

At first, the IDS must catch and screen all traffic inside the system, not only that among itself and a switch. In this manner, it is fitting to utilize reflected ports on the change to empower the IDS to see the entire system.

It is likewise important to put the IDS before the firewall in the system to get maximal discovery of interruption, as some nosy bundles might be sifted through by the firewall.

The issue is that present NIDS are tuned explicitly to distinguish realized administration level system assaults. Endeavours to grow past this restricted domain normally “results in an unsatisfactory dimension of false positives, In the meantime, enough Data exists or could be gathered to permit organize directors to distinguish these arrangement infringement. Sadly, the Data is so voluminous, and the investigation procedure so tedious, that the overseers don't have the assets to experience everything and locate the applicable knowledge”. At the end of the day, arrange chairmen don't have the assets to proactively break down

the Data for strategy infringement, “particularly within the sight of a high number of false positives that reason them to squander their restricted assets”.

Given the idea of this issue, the normal arrangement is Data-mining/AI in a disconnected situation. Such a methodology would add extra profundity to the directors safeguards, and enables them to all the more precisely figure out what the dangers against their system are however the utilization of various strategies on Data from numerous sources. Henceforth, movement that it isn't proficient to recognize in close continuous in an “online NID, either because of the measure of express that would should be held, or the measure of computational assets that would should be used in a constrained time window, can be all the more effectively distinguished”. A several instances of what such a framework could distinguish, that online NIDS can't recognize successfully, incorporate particular kinds of malignant movement, for example, low and moderate outputs, a gradually spreading worm, strange action of a client dependent on some new example of action (instead of a solitary association or modest number of organizations, which will undoubtedly create various false positives), or even new types of assaults that online sensors are not tuned for.

III. CONCLUSION

Due to the urgent demand for effective IDS in network security, researchers are striving to identify improved approaches. This work illustrates how the KDD dataset is very useful for testing different classifiers. The work concentrates on KDD preprocess phase to prepare fair experiments and fully randomized independent test data. Among the classification techniques (J48, MLP and Bayes Network), the J48 classifier has achieved the highest accuracy rate for detecting and classifying all KDD dataset attack types (DOS, R2L, U2R, and PROBE). KDD dataset has attributes and all of them have been recorded, but as part of future work more classifiers will be tested as well as the feature selection to see the most important features.

REFERENCES

[1] Ayres, I (2008). Super Crunchers. New York: Bantam Books.

[2] Bailey-Kellog, C. Ramakrishnan, N. and Marathe, M. (2017) Spatial Data Mining to Support Pandemic Preparedness. SIGKDD Explorations (8) 1, 80-82.

[3] Cao, X., Maloney, K.B. and Brusic, V. (2008). Data mining of cancer vaccine trials: a bird's-eye view. Immunome Research, 4:7. DOI:10.1186/1745-7580-4-7

[4] Cheng, T.H., Wei, C.P., Tseng, V.S. (2006) Feature Selection for Medical Data Mining: Comparisons of Expert Judgment and Automatic Approaches. Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06).

[5] Health Grades, Inc. (2007). The Fourth Annual HealthGrades Patient Safety in American Hospitals Study.

[6] Kou, Y., Lu, C.-T., Sirwongwattana, S., and Huang, Y.-P. (2004). Survey of fraud detection methodologies. In Networking, Sensing and Control, 2004 IEEE International Conference on Networking, Sensing and Control. (2) 749-754.

[7] Nightingale, F (1858). Notes on Matters Affecting the Health, Efficiency and Hospital Administration of the British Army. Shillabeer, A (29 July 2009). Lecture on Data Mining in the Health Care Industry. Carnegie Mellon University Australia.

[8] Shillabeer, A. and Roddick, J (2007). Establishing a Lineage for Medical Knowledge Discovery. ACM International Conference Proceeding Series. (311) 70, 29-37.

[9] Tandoc, E.S (2006). DOH order probe after Rizal hospital tragedy -- Sanitation regulations stressed. Philippine Daily Inquirer, p. A19.

[10] Thangavel, K., Jaganathan, P.P. and Easmi, P.O (2015). Data Mining Approach to Cervical Cancer Patients Analysis Using Clustering Methodology. Asian Journal of Data Technology (5) 4, 413-417.

[11] Tufte, E. (1997). Visual Explanations. Images and Quantities, Evidence and Narrative. Connecticut: Graphics Press.

[12] Wong, W.K., Moore, A., Cooper, G. and Wagner, M (2005). What's Strange About Recent Events (WSARE): An Algorithm for the Early Detection of Disease Outbreaks. Journal of Machine Knowledge Research. 6, 1961- 1998.