*Journal of Advances and Scholarly Researches in Allied Education*
*Vol. XI, Issue No. 22, July-2016, ISSN 2230-7540*

# A Study of Data Mining Techniques towards Network Intrusion Detection

**Mahammad Mastan[1]* Dr. Venkatesh Sharma[2]**

[1] Research Scholar, Sri Venkateshwara University, Uttar Pradesh

[2] Associate Professor, CSE Department, Shri Venkateshwara University Gajraula, Uttar Pradesh

***Abstract – The main purpose of Intrusion Detection Systems(IDS) and Intrusion protection Systems(IPS) for data mining is to discover patterns of program and user activity, and determine what set of events indicate an attack. In the last years, the networking revolution has finally come of age. More than ever before, we see that the Internet is changing computing as we know it. The possibilities and opportunities are limitless; unfortunately, so too are the risks and chances of malicious intrusions. In Network Security, intrusion detection and prevention system is the act of detecting activity or action that attempt to compromise the confidentiality, integrity or availability of a resource. Intrusion prevention techniques, such as user authentication avoiding programming errors, and information protection (e.g., encryption) have been used to protect computer systems is act as first line of defense. We focus on issues related to deploying a data mining-based IDS in a real time of networking environment. To improve accuracy and security, data mining programs are used to analyze audit data and extract features that can distinguish normal activities from intrusions. In this paper presents an intrusion detection system architecture consisting of network sensors, detectors, a data warehouse, and model generation components and we can identify network attack and which type of attack on databases being take place.***

***Keywords: Intrusion Detection Systems, Network Security, Databases, Data Mining.***

----------X----------

## INTRODUCTION

Data mining is additionally for the most part alluded to as knowledge discovery from Data (KDD). The motivation behind Data mining is to mine helpful data from gigantic databases or Data distribution centers. Presently multi day, Data Mining is getting to be normal in human services field in light of the fact that there is a fundamental of operational logical system for identifying unidentified and significant data in wellbeing Data. In industry, Data Mining offers various advantages, for example, recognition of the misrepresentation of data

The effective use of Data mining in profoundly obvious fields like e-business, showcasing and retail has prompted its application in KDD in different ventures and areas.

Data mining algorithms in computer network and demonstrate a vital job in gauge and finding of the ailments. There are various mining applications are build up in the restorative related zones, To get the important and obscure data from the database is the assurance behind the utilization of Data mining. The knowledge discovery is an intuitive procedure, containing by building up a comprehension of the application space, picking and making data collection, pre-preparing, data change.

Data mining is one among the most vital strides in the knowledge discovery process. It tends to be viewed as the core of the KDD procedure. This is the territory, which manages the use of know-how algorithms to get valuable examples from the Data.

A portion of the distinctive methodologies for knowledge utilized in Data mining and as pursue:

**Classification knowledge: -** The data algorithms take a lot of grouped models (preparing set) and use it for preparing the algorithms. With the prepared algorithms, characterization of the test Data happens dependent on the examples and standards separated from the preparation set. Classification can likewise be named as foreseeing a particular class.

**Numeric predication: -** This is a variation of classification knowledge with the special case that as opposed to foreseeing the discrete class the result is a numeric esteem.

**Association knowledge: -** The association and examples between the different characteristics are

removed are from these tenets are created. The tenets and examples are utilized foreseeing the classes or classification of the test Data.

**Clustering: -** The gathering of comparative occurrences in to bunches happens. The difficulties or disadvantages considering this sort of AI is that we need to initially recognize bunches and dole out new times to these groups.

There are a several knowledge methodologies that can be utilized inside each kind of knowledge strategies (E.g. Decision Tree can be considered as a characterization strategy, $K^{th}$ Nearest Neighbor is considered as a bunching system) however paying little heed to the knowledge methodologies, idea is given to the documentation on what is to be discovered and idea depiction is the result delivered by the case after the knowledge method.

Out of these four kinds of knowledge strategies we will be just focusing our work on two, to be specific the classification knowledge and association rules. Various distinctive kinds of order and association systems are referenced in the following part. Classification kind of knowledge is additionally called managed knowledge and bunching is called un-administered knowledge.

***Text mining***

Data can exist in numerous structures, for example, recordings, pictures and text. Data mining can be utilized to remove valuable data from any type of Data. Text mining is the utilization of shrewd algorithms to extricate valuable data from unstructured text.

In text mining the objective is to find obscure data. Accordingly to change over the KDD procedure to outline the text mining process we should supplant every one of the examples of the word Data in Figure 1 by text in every one of the means of the KDD procedure.

Text mining is imperative given that numerous frameworks incorporate databases with traits present in text arrangement. The algorithms in Data mining need not be changed for each sort of Data. Commonly Data must be changed over either to text configuration or to paired arrangement by the compiler before being characterized by the algorithms.

Essentially to Data mining, text mining has numerous applications. Some of them are the accompanying:

Recovering reports: Query preparing assumes an imperative job in effective data recovery. With the assistance of text mining we will most likely successfully produce inquiries that create better outcomes regarding the culmination and the viability of the recovery procedure.

Record ID: The objective of programmed knowledge algorithms is to investigate documents dependent on examples and arrange them in like manner. This objective is practiced by methods for watchwords, which are utilized to distinguish which writer has composed the report and furthermore can be utilized for programmed arrangement of research papers and diaries. This is finished by looking at specialized scientific categorizations, etymology or notwithstanding utilizing the recurrence check strategy (Depending on the recurrence of specific words utilized we can now and again recognize the creator). Prediction or estimating: Based on time arrangement, we can utilize text digging for prediction, which will demonstrate valuable in gauging and observing the progressions that should be made utilizing time touchy examples. One imperative issue in text mining is the presence of copies and irregularities in the Data. For the most part there are situations when text is rehashed or a several properties are available in two distinct scales. There are situations when there are missing traits. With the assistance of pre-processing, we can kill partially the majority of the clamor present in the Data. Data processing results in more prominent effectiveness in running the wise algorithms.

## LITERATURE REVIEW

Liu et al (2006) built up an amusement theoretic system for interruption location in MANET. They display the gate crasher and the protector as a two player Bayesian diversion. They present the Bayesian half breed recognition approaches which screen the system in two distinctive ways, to be specific lightweight and heavyweight observing frameworks. The lightweight observing framework devours less vitality and in this manner it is dependably on, while the heavyweight checking framework utilizes inconsistency based interruption identification framework to construct a typical profile and look at it against the tried information so as to recognize interruption.

Marchang and Datta (2008) drafted two interruption recognition strategies for versatile specially appointed systems, which utilizes community endeavours of hubs in an area to distinguish a malevolent hub in that area. Messages are passed between the hubs and relying upon the messages got, these hubs decide suspected (hubs that are suspected to be vindictive). These speculated hubs are in the long run sent to the screen hub (the initiator of the discovery calculation).

Razak et al (2008) exhibited companion helped interruption location and reaction instruments for portable specially appointed systems through kinship connection.

Otrok et al (2008) planned a bound together structure that can draw out the lifetime of IDS in a group by adjusting the asset utilizations among every one of the hubs. This is accomplished by honestly choosing the most cost-productive hub that handles the

identification procedure. Motivators are given as notorieties to propel hubs in uncovering honestly their expenses of investigation. Notorieties are registered utilizing the outstanding Vickrey, Clarke and Groves (VCG) system where truth-telling is the prevailing technique.

Komninos and Douligeris (2009) proposed a Layered Intrusion Detection Framework (LIDF) to recognize bargained and noxious hubs in a specially appointed system. LIDF comprises of three modules to be specific, accumulation, identification and alarm. These modules work locally in each hub of a system. The accumulation and capacity of review information is performed with the utilization of a parallel tree. The discovery is accomplished with Lagrange inserting polynomials and the alarm is cultivated with straight edge plans.

Al-Roubaiey et al (2010) created versatile affirmation interruption identification for MANET with hub location upgrade. This is an affirmation based plan which can be considered as a mix of plan called TACK and an end - to-end affirmation conspire called AC Knowledge (ACK).

Mohammed et al (2011) portrayed a component configuration based model for secure pioneer decision within the sight of narrow minded hubs. To adjust the asset utilization of the hubs in the system, hubs with the most outstanding assets ought to be chosen as the pioneers. This model has presented a two head decision calculation, to be specific Cluster Dependent Leader Election (CDLE) and Cluster Independent Leader Election (CILE).

Bu et al (2011) proposed a completely conveyed plan of consolidating interruption location and consistent confirmation in MANET. They utilized Dempster–Shafer hypothesis for information combination. In this work multimodal biometrics are sent to mitigate the weaknesses of unimodal biometric frameworks. The biometric confirmation process requires a lot of calculation, the vitality utilization is critical.

Zhao et al (2012) clarified a hazard mindful reaction instrument to adapt to the recognized steering assaults. This methodology utilizes broadened Dempster-Shafer scientific hypothesis with the idea of significance factors. They considered an enhanced connection state steering convention. The hazard mindful reaction system is separated into proof accumulation, chance evaluation, basic leadership and interruption reaction.

**Intrusion Detection Before Data Mining -**When we start the intrusion detection on our organizations network, that time we didn't focus on data mining, but rather on more issues: How alarm generated? How much data would we get? How would we show the data? And what type of data we want to monitor or see? We began to suspect that our system was inadequate for detecting the most dangerous attacks—those performed by adversaries using attacks that are new, stealthy, or both. So we considered data mining with two questions in mind: (Lashari, et. al., 2016)

- Can we develop a way to minimize what the analysts need to look at daily?
- Can data mining help us find attacks that the sensors and analysts did not find?

## KNOWLEDGE DISCOVERY IN DATABASES [KDD] AND DATA MINING:

Traditional strategies (Methods utilized before PCs where brought into medicinal services) utilize manual examination to discover examples or concentrate Data from the database. For instance on account of medicinal services, the wellbeing organizations (E.g. The Center for Disease Control in the US) break down the patterns in maladies and the event rates. This enables wellbeing organizations to avoid potential risk in future in basic leadership and arranging of medicinal services the executives.

The traditional strategy is utilized to examine Data physically for examples for the extraction of knowledge. Take any field like banking, specialist, human services, and showcasing; there will dependably be a Data investigator to work with the Data and breaking down the last outcomes. The examiner demonstrations like an interface between the Data and knowledge. We can, utilizing machine insight help the expert to deliver comparative outcomes or Data from the Data.

When we experience designs inside a database we express the discoveries (examples or standards) as Data mining, data recovery or knowledge extraction, etc. The term Data mining is utilized for the most part by analysts, Data investigators and the administration data frameworks (MIS) (Mohammed, et. al., 2011). The distinction between Data mining and knowledge discovery is that the last is the use of various astute algorithms to remove designs from the Data while Data discovery is the general procedure that is engaged with finding Data from Data. There are different advances, for example, Data pre-processing, Data decision, Data cleaning, and Data representation, which are additionally a piece of the KDD procedure.

## THE KDD PROCESS:

Knowledge discovery is the procedure of consequently creating data formalized in a structure "justifiable" to people (Zhao, et. al., 2012).

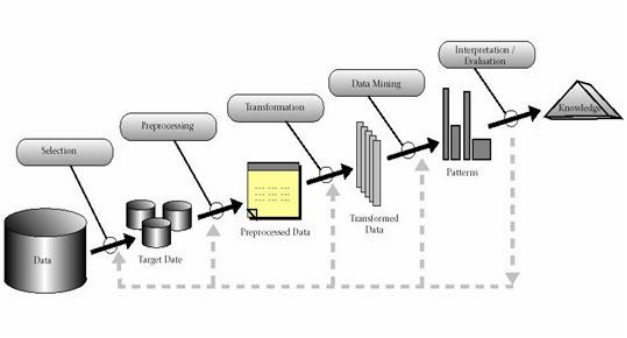


**Figure 1 KDD process**

Three segments are required for the KDD procedure, which are the accompanying:

- A objective is the result we have to discover from examining the Data; Example: what number of individuals with X Y Z indications passed on with malignant growth?

- A database is the place every one of the Data and data about the framework is found. Generally this stage is utilized to know the background data. This data furnished will be connected with the preparation Data or instances gave which is utilized to the following stage. Model, what does this characteristic in the database represent?

- A set of preparing instances, as depicted prior, the framework that is made is robotized, which means the client just need to put in the database and data about what he needs to discover. First the framework ought to be prepared with the goal that it can break down the similitudes between different qualities of the preparation instances. The guidelines acquired can be utilized to anticipate the results in the testing models.

A blueprint of the means that are in Figure 1 will be satisfactory for understanding the ideas required for the KDD procedure. Coming up next are the means included:

**Stage 1:-** The initial step is to predefine our central goal or an objective before finding knowledge. We likewise need to bring up from which database we can get the Data.

**STEP2:-** Consider a situation where we have a large number of Data focuses. We need to choose a subset of the database to play out the required knowledge discovery steps. Decision is the way toward choosing the correct Data from the database on which the apparatuses in Data mining can be utilized to separate data, knowledge and example from the grave crude Data.

**STEP3:-** Data pre-processing and Data cleaning. In this progression we attempt to take out clamor that is available in the Data. Commotion can be characterized as some type of mistake inside the Data. A portion of the devices utilized here can be utilized for filling missing qualities and disposal of copies in the database.

**Stage 4:-** Transformation of Data in this progression can be characterized as diminishing the dimensionality of the Data that is sent for Data mining. Generally there are situations where there are a high number of characteristics in the database for a specific case. With the decrease of dimensionality we increment the productivity of the Data mining venture as for the precision and time usage.

**Stage 5:-** The Data mining step is the significant advance in Data KDD. This is the point at which the cleaned and pre-processed Data is sent into the astute algorithms for characterization, bunching, similitude seek inside the Data, etc. Here we picked the algorithms that are appropriate for finding designs in the Data. A portion of the algorithms give better precision as far as Data discovery than others. Along these lines choosing the correct algorithms can be vital now.

**Stage 6:-** Interpretation. In this progression the mined Data is introduced to the end client in a human-distinguishable arrangement. This includes Data representation, which the client deciphers and comprehends the found knowledge acquired by the algorithms.

## DATA MINING: A PROCESS:

On a very basic level, Data mining is tied in with processing Data and recognizing examples and patterns in that data so we can choose or pass judgment. Data mining standards have been around for a long time; however, with the coming of huge Data, it is much increasingly common.

Enormous Data caused a blast in the utilization of progressively broad Data mining procedures, incompletely in light of the fact that the span of the data is a lot bigger and on the grounds that the data will in general be increasingly differed and broad in its very nature and substance. With vast Data collections, it is never again enough to get generally basic and direct measurements out of the framework. With 30 or 40 million records of itemized client data, realizing that two million of them live in one area isn't sufficient. You need to know whether those two million are a specific age gathering and their normal income with the goal that you can focus on your client needs better.

These business-driven requirements changed basic Data recovery and insights into progressively complex Data mining. The business issue drives an examination of the Data that fabricates a model to portray the data that at last prompts the making of the subsequent report.

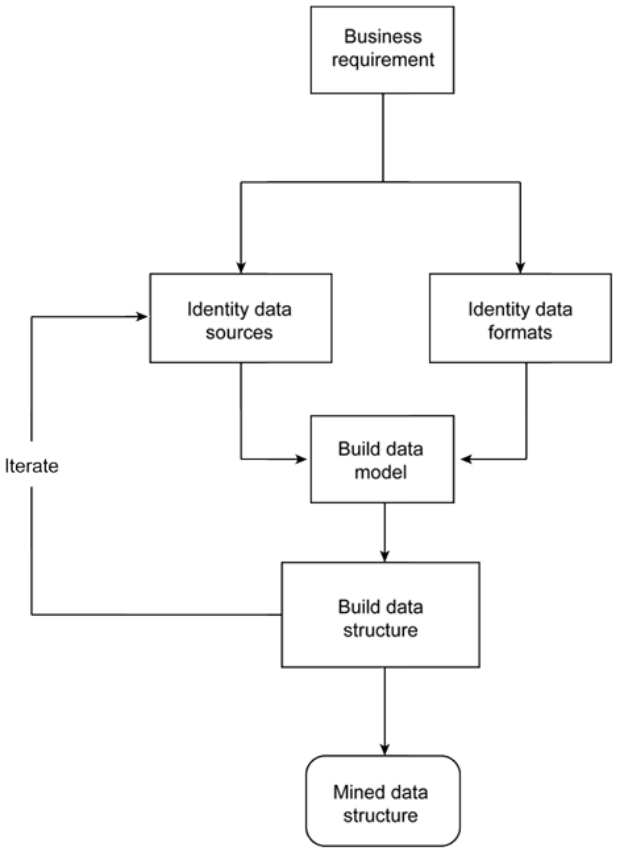


**Figure 2: Process Outline**

"The procedure of information examination, disclosure, and model-building is frequently iterative as you target and recognize the distinctive Data that you can remove. We should likewise see how to relate, guide, partner, and bunch it with other information to deliver the outcome. Recognizing the source information and organizations, and afterward mapping that Data to our given outcome can change after find distinctive components and parts of the information".

## DATA MINING TOOLS

"Data mining isn't about the apparatuses or database programming that you are utilizing. We can perform Data mining with relatively unobtrusive database frameworks and basic instruments, including making and composing your own, or utilizing off the rack programming bundles. Complex Data mining profits by the past experience and algorithms characterized with existing programming and bundles, with specific instruments picking up a more prominent liking or notoriety with various strategies". For instance, IBM SPSS, which has its underlying backgrounds in statistical and overview investigation, can manufacture viable prescient models by taking a gander at past patterns and building precise conjectures. "IBM Info Sphere Warehouse gives Data sourcing, pre-processing, mining, and investigation data in a solitary bundle, which enables you to take data from the source database straight to the last report yield". It is later that the extremely huge Data collections and the bunch and expansive scale Data preparing can permit Data mining to examine and provide details regarding gatherings and connections of Data that are increasingly muddled, Now a completely new scope of apparatuses and frameworks accessible, including joined Data stockpiling and processing frameworks. We can mine Data with an alternate different Data accumulations, including, SQL databases, text Data, key stores and record databases. The databases, for instance, Hadoop, Cassandra, Couch DB, and Couch base Server, store and offer access to Data in order to not facilitate the customary table structure. Specifically, the more adaptable stockpiling configuration of the report database causes an alternate concentration and multifaceted nature regarding preparing the data. "SQL databases impost exacting structures and inflexibility into the pattern, which makes questioning them and breaking down the Data direct from the viewpoint that the arrangement and structure of the data is known". Record databases that have a standard, for example, JSON upholding structure, or documents that have some machine-lucid structure are likewise simpler to process, in spite of the fact that they may include complexities due to the varying and variable structure. For instance, with Hadoop's altogether crude Data processing it tends to be mind boggling to recognize and extricate the substance before you begin to process and correspond it.

## DOCUMENT DATABASES AND MAPREDUCE:

The MapR educe preparing of numerous cutting edge report and NoSQL databases, for example, Hadoop, are intended to adapt to the extremely extensive Data indexes and data that does not generally pursue an unthinkable configuration. When you work with Data mining programming, this idea can be both an advantage and an issue. The primary issue with document based Data is that the unstructured organization may require more processing than you hope to get the data you need. Various records can hold comparative Data. Gathering and orchestrating this data to process it all the more effectively depends upon the arrangement and Map Reduce stages. Inside a Map Reduce-based framework, it is the job of the guide venture to take the source Data and standardize that data into a standard type of yield. This progression can be a generally basic procedure (recognize key fields or Data focuses), or progressively mind boggling (parsse and preparing the data to create the example Data). The mapping procedure creates the institutionalized organization that you can use as your base. Decrease is tied in with abridging or measuring the data and afterward yielding that data in an institutionalized structure that depends on the sums, totals, measurements, or different investigation that you chose for yield.

## METHODOLOGY

In our paper we will propose the following methods for intrusion detection and intrusion prevention system for data miming. Data Mining may be thought of as the most interesting one in accomplishment of intrusion detection and intrusion prevention system. In IDS and IPS, Data Mining used for to discover consistent and useful patterns of system features that describe user behavior. In intrusion detection and intrusion prevention system can be two types.

- Misuse-based system
- Anomaly-based system

**Table 1: comparison among Misuse-based and Anomaly-based**

| Misuse-based | Anomaly-based |
|---|---|
| The attacks uncovered under this are assumed to be true positives. | The normal packets separated under this are assumed to be true negatives. |
| It risks high porosity towards new and undefined attacks. | It risks the chances of normal but undefined packets to be tagged as abnormal data. |
| It has a chance of failure to capture many attacks. | It has a tendency to show greater number of false positives. |

Thus we can introduce INIDS (Integrated NIDS). Not only will INIDS be an integrated system which uses both misuse-based and anomaly based approaches, but it also implements a classification rules again on the data. Data Mining-based intrusion detection systems have demonstrated high accuracy, good generalization to novel types of intrusion, and robust behavior in a changing environment, In Figure 3 we depicted (Peietal.: Data Mining Techniques for Intrusion Detection and Computer Security) (Khanmohammadi and Chun-An, 2016).

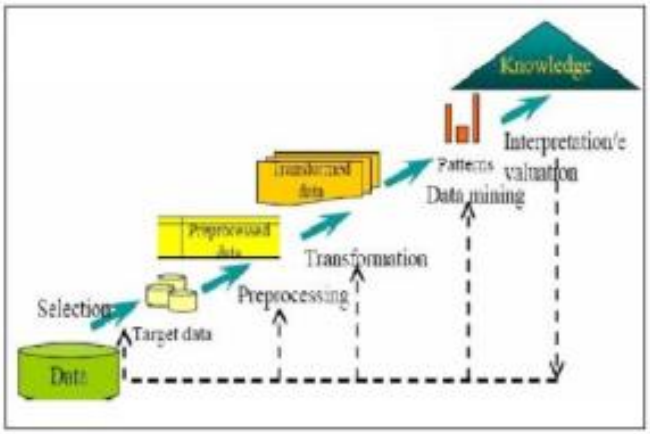


**Fig 3. Data Mining Process Lifecycle**

The intrusion detection and intrusion prevention system is an integrated system which uses both misuse-based and anomaly based approaches. Data mining techniques that are used for intrusion detection and intrusion prevention system are as following,

**a)** **Classification rules:** The classification rules used to discover attacks in a TCP dump. These classification rules used to accurately capture the behavior of intrusions and normal activities for data mining system. The classification rule that we use is the decision tree. Decision Tree: Decision tree induction is the learning of decision trees from class-labeled training tuples. A decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label.

**b)** **Knowledge Discovery in Databases (KDD):** KDD can be defined as "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. Data mining is a particular step in this process in which specific algorithms are applied to extract patterns from data. The KDD process involves a number of steps and is often interactive, iterative and user-driven (Aswal and Ahuja, 2016).

- Getting to know the application domain: trying to understand the data and the discovery task
- Data Mining: includes first deciding what model, for example, summarization, classification, or clustering is to be derived from the data.
- Using the discovered knowledge: includes incorporating the knowledge into a production system, or simply reporting it to interested parties

## CONCLUSIONS:

In this study, identify the type of network attacks take place on the enterprise network or data base systems. The classification rules can be used for intrusion detection and intrusion prevention system (IPS) to classify the network attack and its signature. Many possibilities have been considered, even the incorporation of artificial intelligence. We have shown the ways in which data mining has been known to aid the process of Intrusion Detection and the ways in which the various techniques have been applied for network intrusion detection (IDS) and intrusion prevention system (IPS).

## REFERENCES:

1. Liu, Y. & Man, H. (2006). A Bayesian Game Approach for Intrusion Detection in Wireless

Ad Hoc Networks" Proceedings from the 2006 Workshop on Game Theory for Communications and Networks, Pisa, Italy, pp. 1-12.

2. Marchang, N. & Datta, R. (2008). Collaborative techniques for intrusion detection in mobile ad-hoc networks". Ad Hoc Networks, vol.6, no.4, pp.508-523.

3. Razak, S.A., Furnell, S.M., Clarke, N.L. & Brooke, P.J. (2008). "Friendassisted intrusion detection and response mechanisms for mobile ad hoc networks", Ad Hoc Networks, vol.6, no.7, pp.1151-1167.

4. Otrok, H., Mohammed N., Wang, L., Debbabi, M. & Bhattacharya, P. (2008). A game theoretic intrusion detection model for mobile ad hoc networks", Computer Communications, vol. 31, no. 4, pp. 708-721.

5. Komninos, N. & Douligeris, C. (2009). LIDF: Layered intrusion detection framework for ad-hoc networks", Ad Hoc Networks, vol. 7, no. 1, pp. 171-182.

6. Al-Roubaiey, A., Sheltami, T., Mahmoud, A., Shakshuki, E. & Mouftah, H. (2010). AACK: Adaptive Acknowledgment Intrusion Detection for MANET with Node Detection Enhancement," In proceedings of 24th IEEE International Conference on Advanced Information Networking and Applications, Perth, Australia, pp. 634-640.

7. Mohammed, N., Otrok, H., Wang, L., Debbabi, M. & Bhattacharya, P. (2011). Mechanism Design-Based Secure Leader Election Model for Intrusion Detection in MANET", IEEE Transactions on dependable and Secure Computing, vol. 8, no. 1, pp. 89-103.

8. Zhao, Z., Hu, H., Ahn, G & Wu, R. (2012). Risk-Aware Mitigation for MANET Routing Attacks", IEEE Transactions on Dependable and Secure Computing, vol. 9, no. 2, pp. 250-260.

9. S. A. Lashari, R. Ibrahim, N. Senan, I. T. R. Yanto, and T. Herawan (2016). "Application of Wavelet Denoising Filters in Mammogram Images Classification Using Fuzzy Soft Set," in International Conference on Soft Computing and Data Mining, pp. 529–537.

10. S. A. Lashari, R. Ibrahim, and N. Senan (2014). "Denoising analysis of mammogram images in the wavelet domain using hard and soft thresholding," in Data and Communication Technologies (WICT), 2014 Fourth World Congress on , pp. 353–357.

11. Khanmohammadi, Sina, and Chun-An Chou (2016). "A Gaussian mixture model based discretization algorithm for associative classification of medical data." Expert Systems with Applications 58 pp. 119-129.

12. Aswal, Shobha, and Neelu Jyothi Ahuja (2016). "Experimental analysis of traditional classification algorithms on bio medical dtatasets." In Next Generation Computing Technologies (NGCT), 2016 2nd International Conference on, pp. 566- 568, IEEE 2016

13. Long, Nguyen Cong, Phayung Meesad, and Herwig Unger (2015). "A highly accurate firefly based algorithm for heart disease prediction." Expert Systems with Applications 42, no. 21, pp. 8221-8231

14. 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

**Corresponding Author**

**Mahammad Mastan***

Research Scholar, Sri Venkateshwara University, Uttar Pradesh

mastanmohd@gmail.com