April, 2020 : Vol. 1 Issue 2

American Journal of Advanced Computing

AJAC

SMART

SOCIETY FOR MAKERS, ARTISTS, RESEARCHERS AND TECHNOLOGISTS

6408 ELIZABETH AVENUE SE, AUBURN, WA 98092, USA

U.S. ISSN CENTRE APPROVED

	sequence of alpha & beta thalassemia trait respectively.Here we want to eradicate the thalassemia by systematic method of analysing the defective genetic sequence of the chromosome no 11 & 16 if the conceiving couple are found to be carriers. This is further done amniocentesis by X-rayphotoelectronspectroscopy&D.N.A spectral analysis(that is done by decoding of the graph obtained from spectral analysis using computer algorithms, digital signal processing). DOI: doi.org/10.15864/ajac.1206 Annesha Nayak

Cyber War Fare Defence an Technologies

Sourav Paul

Institute of Engineering and Management, Kolkata

Abstract

The following research paper provides analysis of four (4) **Cyber War Fare Defence and Technologies** topics. These topics include: Virtual Private Network, and Vulnerability Scanning Systems and demonstration of matasploit meterpreter payload. This paper provides basic overview information about each technology, but primarily focuses on analysing each technology within the modern **Cyber War Fare Defence and Technologies** and business context, looking at how it meets business needs while addressing Confidentiality, Integrity and Availability as a Countermeasure that Detects, Corrects and/or Protects. Metasploit , meterpreter ,reverse_tcp demonstration using Infector PC as kali linux Victim PC as WindowsXP(dill injection payload and meterpreter)

I. INTRODUCTION AND OVERVIEW OF APPROACH

This research paper introduces and analyses five (5) Cyber War Fare Defence and Technologies.

Each of the following sections focuses on a specific technology and adheres to the following general format:

o Technology Overview: A high-level introduction to the technology.

o Business Analysis: An evaluation of the usefulness, cost, complexity, and utility of the technology in the modern business environment.

o Security Analysis: The security technology is weighed against the tenets of Confidentiality, Integrity and Availability as well as evaluating its role as a

countermeasure (detect, correct, protect).

The five security technologies addressed in this paper are:

1. Access Control Management

2. Vulnerability Scanning System

3. Virtual Private Network (VPN)

4. The Three Network Vulnerabilities

5. Metasploit about meterpreter

II.ACCESS CONTROL MANAGEMENT

Access control management (ACM) systems pull together identity, authentication and authorization to restrict what resources a user may access and in what manner that access may occur (read, write, execute, modify, etc.). ACM solutions may be based on a number of security models, including Discretionary Access Control (DAC), Mandatory Access Control (MAC), and Role-Based Access Control (RBAC). A standard ACM provides an interface through which a user will self-identify, followed by a mechanism for challenging and confirming that identity, and then a method for granting rights, or access to information, based on the non-repudiated authentication of the user. Access control is at the heart of information security and is the fundamental premise upon which the industry is based1. Without access control management, there would no method through which to provide security for systems and data.

3 National Institute of Standards and Technology, <I>NIST Planning Report 02-1: Economic Impact Assessment of NIST's Role-Based Access Control (RBAC) Program<I> (NIST, 2002, accessed 17 March 2019); available from

A.Business Analysis

Access control management systems provide the foundation for information security within the business environment. Its usefulness is extensive, with the primary functions being to classify data systems according to value and allocate protection mechanisms in accordance with the value of the resource. According to Tipton and Krause, "[the] essence of access control is that permissions are assigned to individuals or system objects, which are authorized to access specific resources."2 The implementation of ACM systems can range in cost from minor to extreme, depending on the value of the resource being 6protected. The underlying security model applied also impacts how expensive and complex the solution may be. ACM solutions are perhaps the most important security technology that can be deployed, ahead of all other countermeasures, because of its inherent purpose to control access to data and systems. The utility of the ACM systems, however, is limitless under the assumption that a business has resources of value that require protecting. Discretionary Access Control systems are very common and are generally cost-effective for most environments. Most operating systems today - ranging from Windows to UNIX to Linux and beyond - make use of a DAC model of access control. Mandatory Access Control systems tend to be more complex and costly in performance and maintenance. MAC systems require a much stronger systematic adherence to the precepts of access control and can thus challenge administrative resources and confound access to data as required by the business. Implementation of MAC requires proper foresight and planning to avoid difficulties in the long term; an effort that is often a costly engineering effort frowned upon by the business. Finally, Role-Based Access Control systems are increasing in popularity and are predicted to saving companies millions of dollars in the coming years.

34 Donald R. Richards, "Biometric Identification," in <I>Information Security Management Handbook, 4th Edition<I>, ed. Harold F. Tipton and Micki Krause (Boca Raton: Auerbach, 2000), p9.

B.Security Analysis

An access control management system has the potential for impacting all three tenets of information security (Confidentiality, Integrity

and Availability). The primary role of an ACM solution is to protect the confidentiality of a resource by restricting access to the resource. Additionally, an ACM solution will control the attributes of the access, such as read, write and execute. For example, in the case of a data file, an ACM system may grant a user read access, but deny access to write or modify the data within the file. Under a DAC model, access controls are managed directly by the resource owner. In a MAC model, the system dictates what level of access may be granted to a resource. Finally, RBAC assigns access based on the rights of a group (or role) within the system. All users who share a given role have the same access. This approach contrasts to DAC where each user may have a unique set of rights. MAC is similar to RBAC in terms of using a role-based approached based on labelling. However, the inner operations of a MAC vary distinctly from an RBAC; discussion of which exceeds the scope of this document.

Access control management systems hinge on the proper identification of subjects trying to access objects. The process of positively identifying a subject is called authentication. The authentication process usually occurs when a subject self-identifies and then responds to a systematic challenge of the identity. This challenge is based on what you know, what you have or who you are. A password is an example of something that you may know, and is currently the most common method of proving identity. A token is an example of something that you have, and biometrics is an example of who you are. Biometrics is a method of identification based on the physical characteristics of a human being, such as a fingerprint, iris scan or retinal scan. Biometrics, though holding significant promise as part of an access control management system, also has significant drawbacks, such as to acceptability to users, reliability and resistance to counterfeiting.4 The future of access control management systems appears to be in the direction of multi factor authentication, oftentimes making use of passwords in combination with tokens or biometrics. Beyond the current trend, it seems likely that passwords will eventually be rendered completely obsolete in favour of some form of token or biometric becoming the first, if not only, form of authentication. Specifically, use of numeric or data tokens is on

the increase and projected to continue gaining in popularity and acceptance. Major international Internet Service Provider America Online has recently announced the availability of numeric tokens for users as a second factor for authentication. Additionally, as public key infrastructure solutions (see Section IX below) mature and gain in prevalence, the use of data tokens will increase in importance. For example, a bank will be able to issue a USB-based data token to a customer. On the data token will be the customer's unique identifier in the form of a digital certificate. This certificate will be managed through a central Certificate Authority and will be used both for authentication and for encrypting and digitally signing communication and transactions. Thus, access control management will not only continue its central role within information security, but it will also grow in scope, adding more extensive capabilities for positively impacting confidentiality and integrity. Additionally, besides protecting resources, it may also include extended capabilities that will allow for easier detection of attacks and possibly even automatic methods for correcting violations of integrity.

1 Ben Rotchke, Access Control Systems & Methodology (SecurityDocs.com, 2004, accessed 16 March 2019); available from

III.VULNERABILITY SCANNING SYSTEMS

Vulnerability scanning is the "automated process of proactively identifying vulnerabilities of computing systems in a network in order to determine if and where a system can be exploited and/or threatened." Vulnerability scanning typically relies on a handful of tools that identify hosts and then proceed to test them for known weaknesses. The automated scanning process should include three high-level steps: receiving authority to scan, determining the scope of the program, and establishing a security baseline (based on the number of vulnerabilities found per number of hosts scanned). Additionally, a good vulnerability scanning program will securely manage the results of the scans and will have a proven plan and process in place for remediation of vulnerabilities that are uncovered. Vulnerability scanning should occur as part of an overall risk management framework, not as a standalone security countermeasure. The most popular vulnerability scanning tool available today is also free, open-source software. Nessus51 has become the de facto tool for vulnerability scanning over the fifteen (15) years, replacing commercial tools like CyberCop Scanner (discontinued), ISS Security Scanner, and eEye Retina. Vulnerability scanning has been around since the late80s or early 90s, pioneered by Dan Farmer, co-author of the COPS52 security tool. Originally, vulnerability scanning was host-based in nature, as COPS and TIGER were, but eventually expanded to include network-based scanning. There are still host-based scanners available, such as the Centre for Internet Security's benchmark security tool53. More often, though, vulnerability scanning today is network-based.

Chapple provides a nice overview of the Nessus scanner and why it’s preferable to its competition:66

"The Nessus tool works a little differently than other scanners. Rather than purporting to offer a single, all-encompassing vulnerability database that gets updated regularly, Nessus supports the Nessus Attack Scripting Language (NASL), which allows security professionals to use a simple language to describe individual attacks. Nessus administrators then simply include the NASL descriptions of all desired vulnerabilities to develop their own customized scans.

Webopedia, vulnerability scanning (Darien: Jupitermedia, undated, accessed 17 March 2019); available from http://www.webopedia.com/TERM/V/vulnerability_scanning.html; Internet. 50 Christopher Cook, Managing Network Vulnerabilities in a DOE/NNSA Environment (Kansas City: DOE, undated, accessed 17 March 2019); available from

A. Business Analysis

vulnerability scanning is a very cheap and useful practice. When conducted regularly and carefully, the use of an automated vulnerability scanning tool can provide considerable information about the overall risk landscape of technologies throughout an enterprise. Vulnerability scanning is particularly important for ensuring that Internet-accessible resources are properly secured before deployment, and to ensure that they remain secure after deployment. Because the most common tools for conducting vulnerability scans is free, open-source software, there is very little reason not to make use of it. Furthermore, the installation and operation of a tool like Nessus does not require much technical acumen. More importantly, the information that can be gathered from the assessment can be invaluable. Operation of a basic vulnerability scanner is not complex. Making matters even better, tools like Nessus are thoroughly

documented on the Internet and can often be found in pre-packaged bootable environments.

B. Security Analysis

Vulnerability scanning can contribute to countermeasures in all three areas of *protect*, *detect* and *correct*. The primary role of the scanning is to detect vulnerabilities in systems, but when used properly it will also contribute to protecting resources from being deployed insecurely and by providing adequate information to allow system administrators to correct vulnerabilities. From the standpoint of Confidentiality, Integrity and Availability, vulnerability scanning most affects the Integrity of systems, though there may be ancillary benefits to Confidentiality and Availability. In detecting and resolving weaknesses in a system, the integrity of the system can be assured. Furthermore, ensuring the integrity of a system will help prevent the system from becoming compromised, resulting in a loss of confidentiality, or from being overly susceptible to attacks that may result in denying the availability of the system or associated application.

Robert Moskowitz, What Is A Virtual Private Network? (Unknown: CMP, undated, accessed 18 March2019); available from; Internet. 48 Elizabeth D. Zwicky and others, Building Internet Firewalls, 2nd Edition (Cambridge: O'Reilly, 2000), p119.

IV.VIRTUAL PRIVATE NETWORKS (VPN)

A Virtual Private Network (VPN) is a private communications network that makes use of public networks, oftentimes for communication between different organizations. A VPN is not inherently secure, though in its most common incarnation it does utilize encryption to ensure the confidentiality of data transmitted. The VPN is often seen as a cheaper solution for deploying a private network than private leased-lines. They often serve to protect and ensure the integrity of communications and may also protect the confidentiality of those communications when utilizing encryption. Aside from the cost factor, VPNs have two main advantages: they may provide overall encryption for communications and they allow the use of protocols that are otherwise difficult to secure. In contrast, Zwickey sites the two main disadvantages of VPNs being the reliance on "dangerous" public networks and extending the network that is being protected. There are three types of VPNs available today: dedicated, SSL and opportunistic.

Dedicated VPNs, either in a gateway-to-gateway or client-to-gateway configuration, appear to currently be the most prominent deployment. However, SSL VPNs are increasing in popularity, serving as a lightweight, platform-independent client-to-gateway protection mechanism. Additionally, the concept of opportunistic encryption, as used with VPNs, was first posited in 2001 by the FreeS/WAN project, who's mission was to provide free standards-based VPN software under an open-source initiative. The concept of opportunistic encryption (OE) hinged on the notion that a VPN did not need to be in an "up" state at all times, but rather only needed to be activated when communication was occurring. Thus, gateways across the Internet could be configured to support encryption on an as-needed basis and would only have to setup the VPN when a connection from/through an OE-aware gateway was initiated. This model is similar to the traditional use of SSL on the Internet, except that instead of simply encrypting the traffic at the application layer, the encryption was actually occurring at the network and/or transport layer, and all happening transparent to the end-user. The goal of implementing opportunistic

encryption within free IPSEC-based VPNs was to transparently encrypt all Internet traffic.

Most virtual private networks today make use of IPSEC encryption. IPSEC provides network-level security for the Internet Protocol (IP) and is an extension of the original IPv4 standard. IPSEC makes use of the management and security protocol ISAKMP/Oakley and has the benefit of protecting against man-in-the-middle attacks during connection setup. IPSEC includes a number of other features, such as being usable by tunnelling protocols.

About.com has several links on VPNs that may be worth reviewing.

Wikipedia, Virtual private network (Wikipedia, 2019, accessed 16 March 2019); available from; Internet. 41 Elizabeth D. Zwicky and others, Building Internet Firewalls, 2nd Edition (Cambridge: O'Reilly, 2000), p104. 42 Robert Moskowitz, What Is A Virtual Private Network? (Unknown: CMP, undated, accessed 16 March 2019); available from; Internet. 43 Elizabeth D. Zwicky and others, Building Internet Firewalls, 2nd Edition (Cambridge: O'Reilly, 2000), p119. 44 Elizabeth D. Zwicky and others, Building Internet Firewalls, 2nd Edition (Cambridge: O'Reilly, 2000), p120. 45 Elizabeth D. Zwicky and others, Building Internet Firewalls, 2nd Edition (Cambridge: O'Reilly, 2000), p121.

A. Business Analysis

Virtual private networks have a legitimate use in the business environment, especially when used in a secure manner, leveraging available encryption options. Given the growing prevalence and availability of cheap Internet access, a VPN can be used to securely and reliably replace more expensive leased lines. This replacement is particularly nice in environments where the data being transmitted is sensitive, but where interruption of connectivity will not represent a major disruption to the business. Many hardware and software solutions are available today, with costs ranging from free (*FreeS*/WAN) to expensive (dedicated hardware-based solutions targeting high throughput). Most inexpensive networking equipment, such as the *Linksys* and *Netgear* lines of home user security devices, now support IPSEC-based VPNs.

46 Henry Spencer and D. Hugh Redelmeier, Opportunistic Encryption (Unknown: Freeswan.org, 2001, access16 March 2019); available from

B. Security Analysis

The basic goal of a Virtual Private Network is to ensure the integrity of the connection and communications. When encryption is added, the goal of preserving confidentiality may also be achieved. One downside to VPNs is that they tend to be built on complex systems and are prone to easy disruption, reducing the overall availability of data and communications. From the perspective of countermeasures, the VPN primarily serves to protect data, though it may also dynamically correct. If logging is enabled and monitored, then attacks against the VPN may also result in meeting the need of detection, though that would be ancillary.

V. NETWORK VERNABILITY

A.Injection vulnerabilities

Injection vulnerabilities occur every time an application sends untrusted data to an interpreter. Injection flaws are very common and affect a wide range of solutions. The most popular injection vulnerabilities affect SQL, LDAP, XPath, XML parsers and program arguments. As explained in the OWASP “Top 10” guide, the injection flaws are quite easy to discover by analysing the code, but frequently hard to find during testing sessions when

systems are already deployed in production environments. The possible consequences of a cyber-attack that exploits an Injection flaw are data loss and consequent exposure of sensitive data, lack of accountability, or denial of access. An attacker could run an Injection attack to completely compromise the target system and gain control on it. The business impact of an Injection attack could be dramatic, especially when hacker compromise legacy systems and access internal data.

SQL injection vulnerabilities are among most exploited flaws, despite the high level of awareness on the various techniques of hacking that exploit this category of bugs the impact of such attacks is very serious.

A study released by the **Ponemon Institute** in October 2014 titled "*The SQL Injection Threat Study*" investigated on the reply of organizations to the SQL injection threat.
The study revealed that despite about one-third believing that their organization has the necessary technology to detect and mitigate the cyber threat, the success rate of SQL injection attacks is too high.

Injection vulnerabilities could affect various software and their impact depends on the level of diffusion of the vulnerable application.

A classic example of the possible effect of the presence of injection flaws is the critical vulnerability dubbed Bash Bug affecting the Linux and UNIX command-line shell. The flaw, coded as ***CVE-2014-6271***, is remotely exploitable and affects Linux and Unix command-line shell potentially exposing to risk of cyber-attacks websites, servers, PCs, OS X Macs, various ***home routers***, and many other devices.
The vulnerability has existed for several decades and it is related to the way bash handles specially formatted environment variables, namely exported shell functions. To run an arbitrary code on affected systems it is necessary to assign a function to a variable, trailing code in the function definition will be executed.

The critical Bash Bug vulnerability, also dubbed Shellshock, affects versions GNU Bash versions ranging from 1.14 through 4.3, a threat actor could exploit it to execute shell commands remotely on a targeted machine using specifically crafted variables.

Such kind of vulnerabilities could have a dramatic effect on a large scale, let's think for example to the dangers for the ***Internet-of-things*** devices like smart meters, routers, web cameras and any other device that runs software affected by this category of flaws.

B.Buffer Overflows

A buffer overflow vulnerability condition exists when an application attempts to put more data in a buffer than it can hold. Writing outside the space assigned to buffer allows an attacker to overwrite the content of adjacent memory blocks causing data corruption, crash the program, or the execution of an arbitrary malicious code.

Buffer overflow attacks against are quite common and very hard to discover, but respect the injection attacks they are more difficult to exploit. The attacker needs to know the memory management of the targeted application, the buffers it uses, and the way to alter their content to run the attack.In a classic attack scenario, the attacker sends data to an application that store it in an undersized stack buffer, causing the overwriting of information on the call stack, including the function's return pointer. In this way, the attacker is able to run its own malicious code once a legitimate function is completed and the control is transferred to the exploit code contained in the attacker's data.There are several types of buffer overflow; most popular are the Heap buffer overflow and the Format string attack. Buffer overflow attacks are particularly dangerous; they can target desktop applications, web servers, and web applications.An attacker can exploit a buffer overflow to target a web application and execute an arbitrary code. He can corrupt the execution stack of a web application by sending specifically crafted data.Buffer overflows affecting widely used server products represent a significant risk to

users of these applications, in the last years, many buffer overflow vulnerabilities were discovered in a number of SCADA components.Considering that the number of cyber-attacks against SCADA is increasing even more it is likely that these buffer overflow vulnerabilities will be exploited with increasing frequency.A number of crimeware kit could be sold in the underground ecosystem to attack this particular category of targets causing serious damages.

C.Sensitive Data Exposure

Sensitive data exposure occurs every time a threat actor gains access to the user sensitive data.Data could be stored (at rest) in the system or transmitted between two entities (i.e. servers, web browsers), in every case a sensitive data exposure flaw occurs when sensitive data lack of sufficient protection.

Sensitive data exposure refers the access to data at rest, in transit, included in backups and user browsing data.The attacker has several options such as the hack of data storage, for example by using a malware-based attack, intercept data between a server and the browser with a Man-In-The-Middle attack, or by tricking a web application to do several things like changing the content of a cart in an e-commerce application, or elevating privileges.The principal sensitive data exposure flaw is the lack of encryption for sensitive data, but even if encryption mechanisms are implemented, other events concur to the exposure of information. The adoption of weak key generation and management, and weak algorithm usage is very common in many industries and applications.A number of incidents recently occurred have demonstrated the critic of this category of flaw, let's think to the wrong implementation of encryption algorithms and the lack of encryption for mobile and cloud solutions.In September 2014, the CERT Coordination Centre at Carnegie Mellon University (CERT/CC) published the results of the tests conducted by its experts on popular Android applications that fail to properly validate SSL certificates.The failure of the certificate pinning procedure exposes users to the risk of ***MitM attacks*** and consequent theft of sensitive information.The CERT confirmed that the problems is widespread, the circumstance was confirmed by another study conducted by security experts at FireEye that *evaluated* the level of security offered by 1,000 of the most popular free apps offered on Google Play.

VI. Metasploit: About Meterpreter

Meterpreter is a tool that is packaged together with the metasploit framework. The features of meterpreter are:

1. Does not create any files on the harddisk, it resides in memory and attaches itself to a process.
2. client-server communication is in the form of type-length-value (TLV) format.

3. client-server communication between attacker machine and victim machine is encrypted.
4. It provides a platform to write extensions.

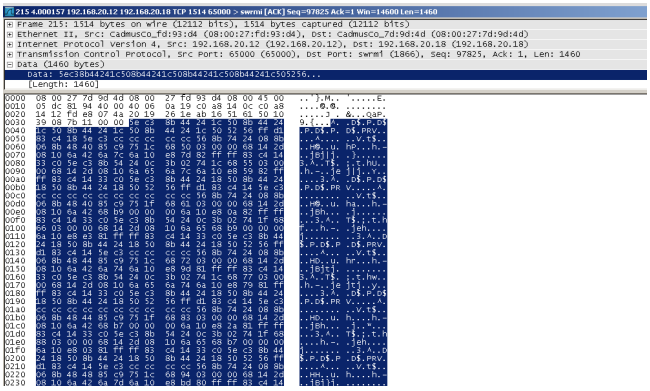


Fig. 1:*Data is encrypted. 192.168.20.12 is the attacker and 192.168.20.18 is the victim.*

How it works:

Step 1: Apply exploit and 1st stage payload (such as reverse tcp binding) to the victim machine.

Step 2: Victim machine connects (using reverse tcp binding) back to attacker's machine.

Step 3: Meterpreter on the attacker's machine sends the 2nd stage payload that does DLL injection.

Step 4: Meterpreter on the attacker machine sends server DLL to the victim machine.

Step 5: Client-server communication establishes.

Demonstration

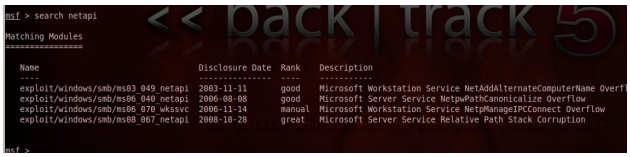


Fig. 2: The victim machine is a Windows XP which is vulnerable to netapi exploit. Choose the exploit with the great ranking.

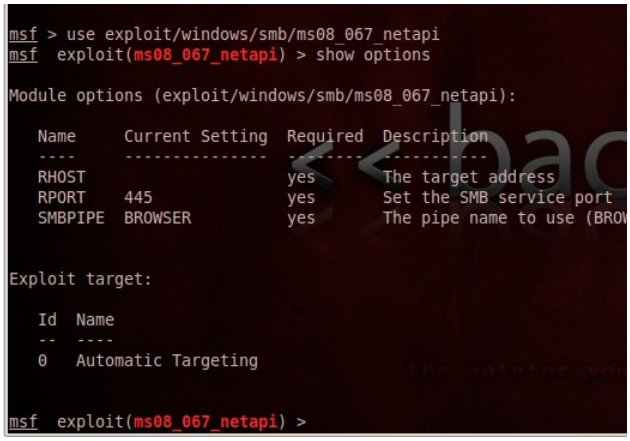


Fig. 3: Use the ms08_067 exploit. The configurable options are shown.

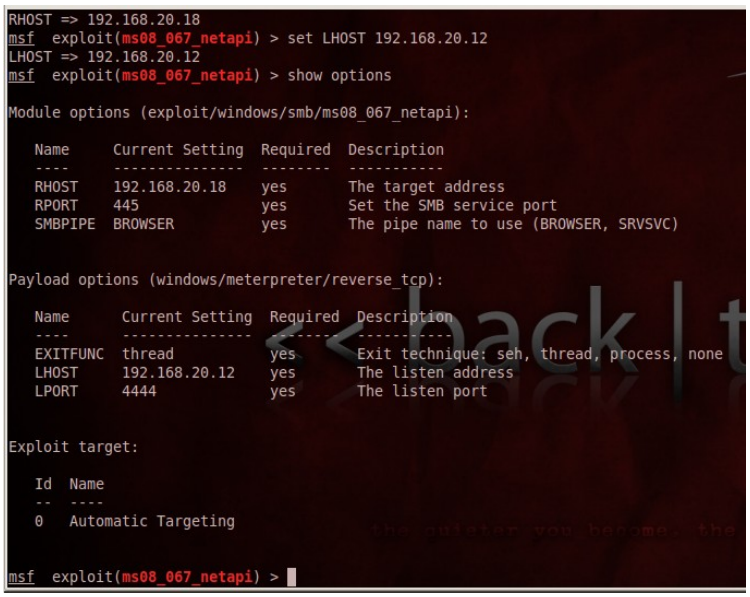
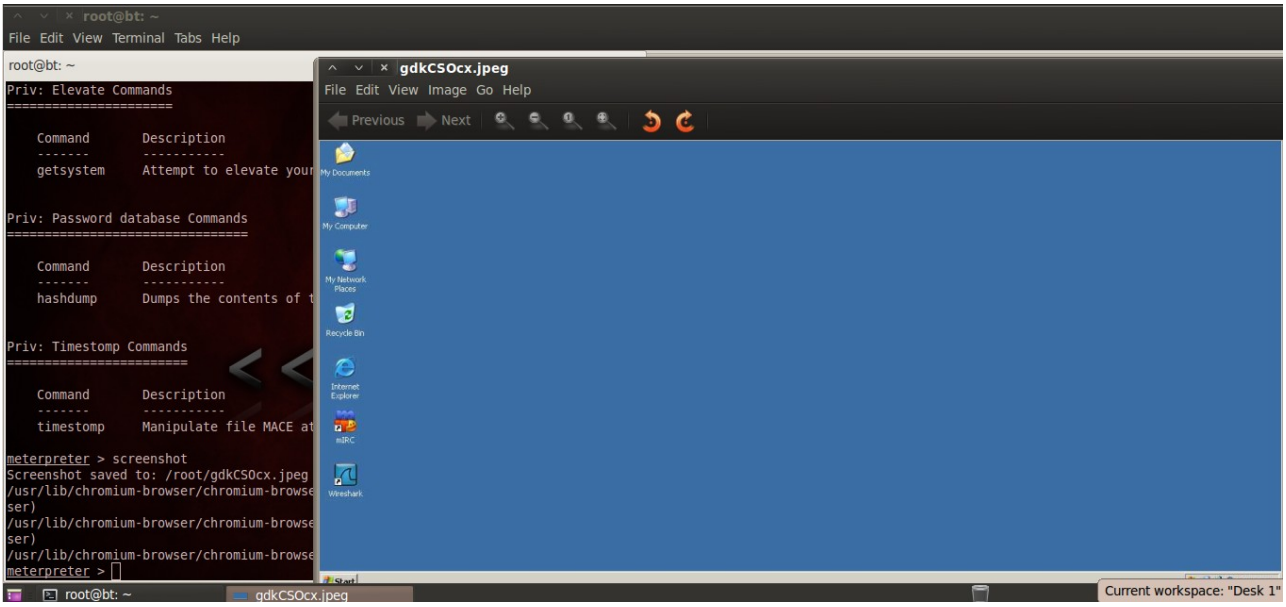


Fig. 6:remote host is the victim, local host is the attacker.

Fig. 4:*Start the exploit. Meterpreter sent 752KB of payload to the victim. No errors, and a meterpreter prompt appeared mean the exploit was successful.*

For a list of meterpreter commands use help or ?

Fig. 5:*As shown, meterpreter has attached itself to svchost.exe*



meterpreter can attach itself to another process using the migrate command. For this demo, attacker migrated meterpreter from svchost.exe to winlogon.exe which is pid 540.

For some reason the keyscan_dump was not working….Meterpreter can do keylogging => http://www.offensive-security.com/metasploit-unleashed/Keylogging

Fig. 7: To know your user privilege level.

Obtain a screenshot of victim's machine

Reference

1. http://csrc.nist.gov/rbac/rbac-impact-summary.doc; Internet.

2. http://www.securitydocs.com/go/69; Internet

3.http://cio.doe.gov/Conferences/Security/Presentations/CookC.pps; Internet.

4.http://compnetworking.about.com/od/vpn/ 40

5.http://en.wikipedia.org/wiki/Virtual_private_network

6.http://www.networkcomputing.com/905/905colmoskowitz.html

7.https://resources.infosecinstitute.com/the-top-five-cyber-security- vulnerabilities-in- terms-of-potential-for-catastrophic-damage/

8.https://resources.infosecinstitute.com/the-top-five-cyber-security-vulnerabilities-in-terms-of-potential-for-catastrophic-damage/

9.http://www.offensive-security.com/metasploit-unleashed/Keylogging

10.https://resources.infosecinstitute.com/the-top-five-cyber-security-vulnerabilities-in-terms-of-potential-for-catastrophic-damage/

11.http://www.freeswan.org/freeswan_trees/freeswan1.91/doc/opportunism.spec; Internet

12.http://www.offensive-security.com/metasploit-unleashed/Keylogging

13.https://resources.infosecinstitute.com/the-top-five-cyber-security-vulnerabilities-in-terms-of-potential-for-catastrophic-damage

14.http://mandeeplubana.blogspot.com/2011/02/meterpreter-is-advanced-dynamically.html

15.http://www.securitytube.net/video/801

16.http://www.offensive-security.com/metasploit-unleashed/Main_Page51

17.http://www.nessus.org/ 52

18.http://www.fish.com/cops/overview.html 53

19.http://www.cisecurity.com/ 54 Mike Chapple, Vulnerability scanning with Nessus (Unknown: TechTarget.com, 2003, accessed 18March2019); available from

20.http://searchsecurity.techtarget.com/tip/0,289483,sid14_gci938271,00.html

21.http://compnetworking.about.com/od/vpn/ 40 Wikipedia, Virtual private network (Wikipedia, 2019, accessed 16 March 2019); available from

22.http://en.wikipedia.org/wiki/Virtual_private_network; Internet. 41 Elizabeth D. Zwicky and others, Building Internet Firewalls, 2nd Edition (Cambridge: O'Reilly, 2000), p104. 42 Robert Moskowitz, What Is A Virtual Private Network? (Unknown: CMP, undated, accessed 16 March 2019); available from

23.http://www.networkcomputing.com/905/905colmoskowitz.html; Internet. 43 Elizabeth D. Zwicky and others, Building Internet Firewalls, 2nd Edition (Cambridge: O'Reilly, 2000), p119.

24. 44 Elizabeth D. Zwicky and others, Building Internet Firewalls, 2nd Edition (Cambridge: O'Reilly, 2000), p120. 45

25. 46 Henry Spencer and D. Hugh Redelmeier, Opportunistic Encryption (Unknown: Freeswan.org, 2001, access16 March 2019); available from http://www.freeswan.org/freeswan_trees/freeswan1.91/doc/opportunism.spec; Internet

26.http://www.offensive-security.com/metasploit-unleashed/Keylogging

27.https://resources.infosecinstitute.com/the-top-five-cyber-security-vulnerabilities-in-terms-of-potential-for-catastrophic-damage/

28.http://www.networkcomputing.com/905/905colmoskowitz.html

29.http://www.freeswan.org/freeswan_trees/freeswan1.91/doc/opportunism.spec

A Shapley Value based Approach to Market Basket Analysis

Rahul Lakhotia
Mindtree
rahullakhotia9@gmail.com
Kolkata, India

Pratibha Goenka
Self employed
pratibhagoenka12@gmail.com
Kolkata, India

***Abstract*— Market Basket Analysis is an interesting concept wherein a set of historical purchase data is studied extensively and data-mining techniques are applied to predict a user's purchase behavior. In the resent times, every e-commerce giant, like Amazon, Flip-kart etc. are trying to increase their sales by this technique. It has been found in study that when a user is purchasing an item then he/she might be interested in purchasing other items too. So, by analyzing this purchase behavior if we can recommend other products at the time of purchasing then that could increase the sale and profit as well. Analyzing the correlation among the products by studying the purchase pattern and then finding the associated products is the main objective of Market Basket Analysis. Several algorithms have been developed to find the association among the products. But most of the traditional algorithms are based on support value of the product. They do not consider another important factor i.e. the marginal contribution of the product. Considering this into account in this paper we are going to propose a Shapley Value based game theoretic approach for market basket analysis. Shapley Value is a well-known solution concept in cooperative game theory. It gives the measure of marginal contribution of a player in a cooperative game.**

Keywords— Market Basket Analysis, Shapley Value, Apriori Algorithm, Support, Association-Combination Algorithm, Association rule mining.

I. INTRODUCTION

Market Basket Analysis is basically a data mining method which works by extracting the co-occurrences from a store's transactional data to analyze the customer's buying behavior. Whenever a person purchases items from a supermarket, the details of his/her purchases are recorded in the store's database. This huge data can then later be thoroughly analyzed to determine a lot of things which can help the store to increase its sales like a customer's purchasing pattern. It can also help the store to take other useful decisions like cross selling, upselling, which items to stock up, which items to stock down, the arrangement of the products in the store etc. This is done through a process called Association Rule Mining (ARM) [1]. This technique identifies the relationship between the items sold in the past by analyzing the already stored large set of database. Association rule mining is used in various industries and not just in e-commerce. Some of the examples include purchase decisions in supermarkets, such as mail order, fraud detection of credit card and telemarketing production. Companies, lately, have been investing a significant amount in collecting purchase data of customers to extract vital information from product feature databases to gain competitive advantages. Companies know that they must be aware of the needs of customers and must act upon their customer's expectations. But this has somewhat proved to be a challenge for many firms. Market Basket Analysis has been thus used widely to discover all the associations between the products and to understand which all items to be put together. Companies can then experiment with their sales by putting lucrative and innovative offers by understanding the psychology of a customer's buying behaviors which have proved to be an important thing to do in the competitive world.

II. MOTIVATION

Data Mining provides immense opportunities in the market sector. In the competitive world, it has become very important for companies to understand the behavior of customers and the challenge is to keep innovating, sustain in this competitive world and live up to the expectations of the modern customer. The main challenge for companies are they have been investing a large amount of resources in extracting significant information from their vast database to come up with some fruitful strategy which will reflect clearly in their sales with a good growth of margin.

The most important goal here in this data mining technique (Market Basket Analysis) is to find out most of the hidden knowledge from the database and use different algorithms that have been proposed for the same. The problem here is all the rules are not equally useful for a particular data set – only a small fraction of the developed rules are interesting to any user. Hence numerous concepts like confidence [5], support [5], and lift have been used to determine which all the most interesting rules among all. But in most of the cases the state-of-the-art algorithms focus on the support and confidence parameter to determine the association rule among the items in Market Basket Analysis. The traditional approaches do not consider the marginal contribution of an item while determining association rules. In this paper we have proposed a Shapley Value based approach for Market Basket Analysis to consider the marginal contribution of each item.

III. THEORETICAL CONCEPTS

In this paper we have proposed a Shapley Value based Apriori algorithm. There are some theoretical foundations behind this approach. These theoretical concepts are discussed in this section.

A. Support

Support is defined as the combined percentage of any two items: identify the combination of the item which fulfills the minimum requirement of support value. Support value of an item is obtained by using the formula as follows:

S (A) = Amount of transaction A (1) / Total transaction

The formula of support value of two items is:

S (A and B) = Amount of transaction of A& B (2) / Total Transaction

B. *Confidence*

The frequencies of the item Y appearing in the transaction which contains X are called the Confidence. After all of the system of high frequency is found, then rules are to be found.

Conf= P(Y | X) = Amount A and B (3) / Amount A

C. *Apriori Algorithm*

Apriori algorithm [3] [6] is a level-wise technique which counts transactions. This algorithm uses an iterative method called level-wise search, in which n-item sets are used to explore (n+1) - item sets. The Apriori algorithm is so named since it uses prior knowledge of frequent *item s*et properties stored in the large collection of dataset to derive results [2] and this property is known as the Apriori property.

Apriori property states that if an item X is joined with item Y,

Support (XUY) =min (Support(X), Support(Y))

An *item set* is any subset of all the items in the database of transactions.

D. *Shapley Value*

In cooperative game theory, Shapley value [4] is a very popular solution concept which gives a unique allocation to a group of players in a coalitional game. Shapley value gives unique expected payoff for individual in a Transferrable Utility game [7]. The physical significance of Shapley value is that it provides the marginal contribu-tion of a player among the group of players.

IV. PROPOSED WORK

After going through the various research papers that have been mentioned above, we have come up with an algorithm. In this algorithm, first the values of all distinct items from the given dataset are extracted. After that Support and Shapley value [4][9] for each of the items is calculated. Then we are calculating the influence of that item by the following formula

Influence of the item = Support(item)*0.6 + ShapleyValue(item)*0.4

Algorithm: Shapley Value based item set detection
Input: List of items of any transaction data
A real value p between 1 to 100
Cardinality of the expected itemset C
Output: A item set with cardinality C

1. For each itemset calculate Support and Shapley Value
2. For each itemset compute its marginal importance (marimp) marimp = Support(itemset)*0.6 + ShapleyValue(itemset)*0.4
3. Create a list of sorted the items in non-increasing order of their marginal importance
4. Calculate k= total number of elements in list * (p/100)
5. Take top k items from the sorted list
6. Increase the size of item set by one, by taking all possible combination among the k items
7. Repeat steps 1 to 6 until the cardinality of the itemset is C
8. Return itemset of cardinality C

V. EXPERIMENTAL RESULTS

We have taken grocery dataset [8] for our experiments. The dataset contains 9835 transaction data for grocery shop. We applied our proposed algorithm on this dataset considering we cardinality of result set is 4. The results are given below.

Fig. 1. These are the set of top 20 single items

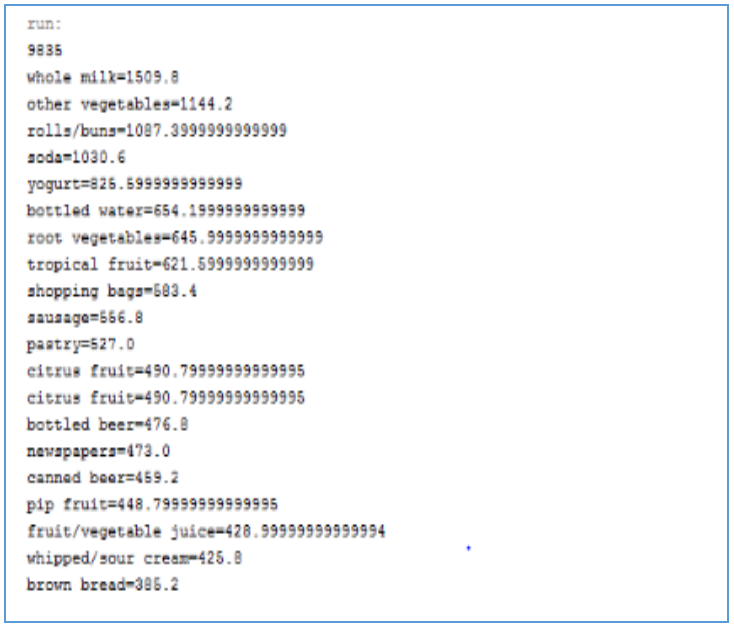


Fig. 2. These are the set of top twenty items of 2 items

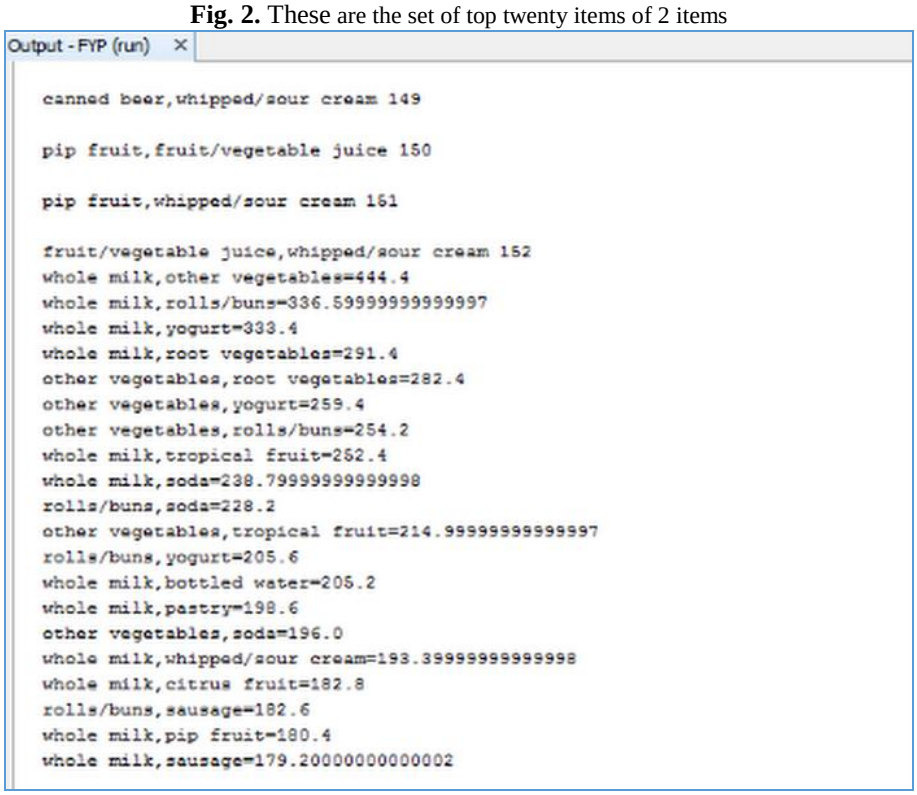


Fig. 3. These are the set of top twenty items of 3 items

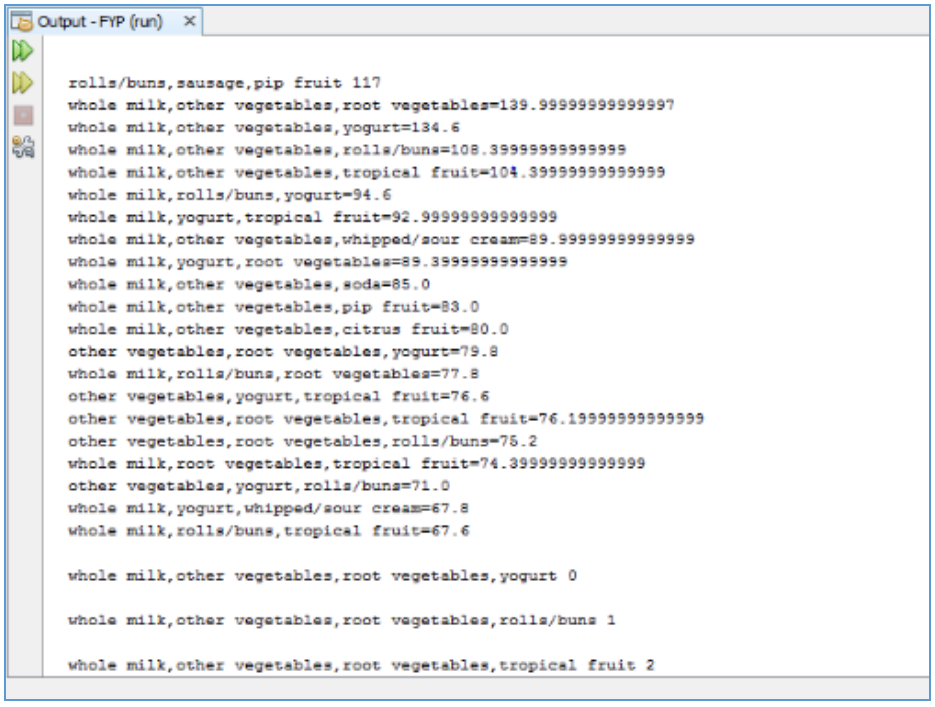


Fig. 4. These are the set of top twenty items of 4 items

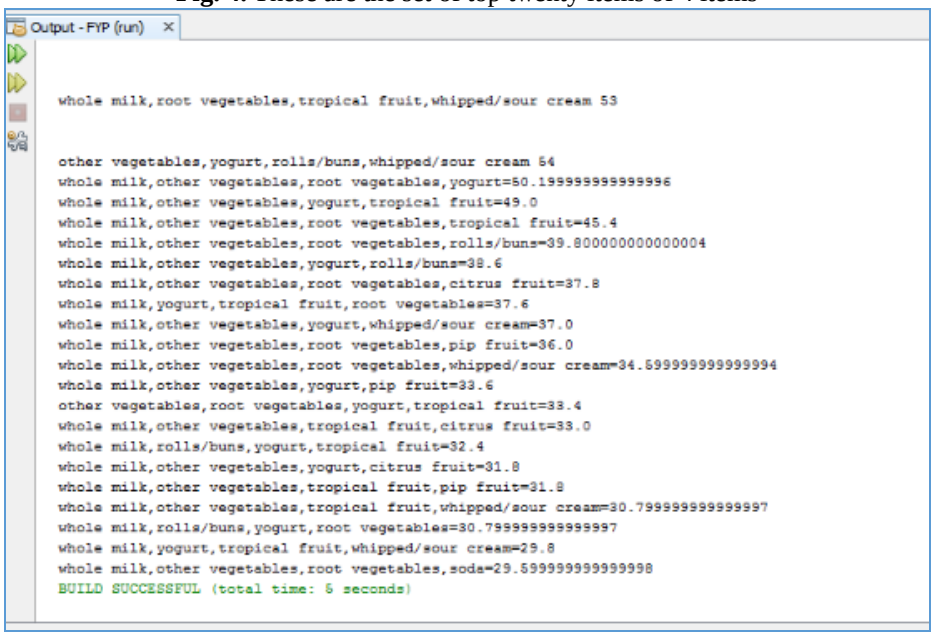
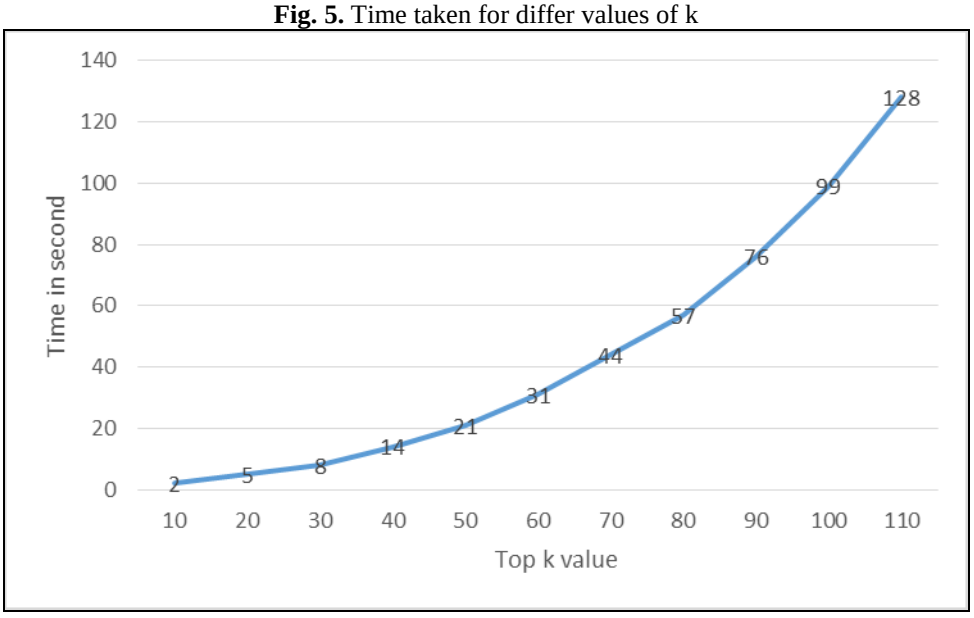


Fig. 5. Time taken for differ values of k



We have performed the experiments for different values of k (Refer Algortihm in Table-1). The time taken for different values of k is shown in Fig-5.

VI. Conclusion

In this paper we have proposed a Shapley value [10] based Market Basket Analysis. Shapley value is a solution concept of cooperative game theory. As game theoretical approach is much more realistic than any probabilistic approach, so our proposed approach is very intuitive. As we are finding the item set up to a certain cardinality, so our approach can be used to find the most frequent item set of any cardinality. We conclude this paper by stating that our proposed approach is not only giving good result like other

References

[1] Feature Selection Based on the Shapley Value by Shay Cohen and Eytan Ruppin, School of Computer Sciences, Tel-Aviv University, Tel-Aviv 69978, Israel, cshay,ruppin_@post.tau.ac.il

[2] Selection Method Under Population Data by Ramesh Prasad Aharwal, Asstt. Prof., Department of Mathematics, Govt. P.G. College, Damoh, M.P. (India)

[3] Market Basket Analysis with Network of Products; Supervisor: Prof. Salvatore Orlando; Candidate: Nikhil Verma, Matriculation number – 855183. Academic Years: 2016/2017

[4] A Comparative Study on Market Basket Analysis and Apriori Association Technique by Warnia Nengsih, Department of Computer, Politeknik Caltex Riau Riau –Indonesia, email: warnia@pcr.ac.id

[5] Market Basket Analysis, A project report Submitted to Department of Computer Science and Information Technology, DWIT College by Sanjeev Mainali, August 2016

[6] Shweta, Ms, and Dr Kanwal Garg. "Mining efficient association rules through apriori algorithm using attributes and comparative analysis of various association rule algorithms." International Journal of Advanced Research in Computer Science and Software Engineering 3.6 (2013).

[7] Nowak, Andrzej S., and Tadeusz Radzik. "A solidarity value forn-person transferable utility games." International Journal of Game Theory 23.1 (1994): 43-48.

[8] Marafi, Salem. "Market Basket Analysis with R." Salem Marafi, www.salemmarafi.com/code/market-basket-analysis-with-r/.

[9] Aadithya, Karthik V., et al. "Efficient computation of the shapley value for centrality in networks." *International Workshop on Internet and Network Economics*. Springer, Berlin, Heidelberg, 2010.

[10] Dragan, Irinel. "The least square values and the Shapley value for cooperative TU games." *Top* 14.1 (2006): 61-73.

A Study on Pollution Prediction and Prevention using IoT and Machine Learning

Shamik Kumar Roy
Academy of Technology
Email: shamik.kumarroy@aot.edu.in

Sahitya Mondal
Academy of Technology
Email: sahityamondal@hotmail.com

ABSTRACT
Climate change and Environmental Hazards has been burning issues all around the world. Air Pollution is a major contribution to the Environmental Pollution. Using Big Data and machine learning algorithm to formulate a solution to this burning global issue with an idea that applies techniques of IoT (Internet of Things) and Data Analytics to predict and prevent air pollution substantially. In this paper the main concern is to judge different works which are related to the air pollution and prevention mechanism which will definitely help the researchers for this domain.

Keywords
Microcontroller, Gas detecting sensors, GSM module, cloud, GPRS, Big Data, Machine learning, MLP NN and ELM.

1. INTRODUCTION
Air pollution is extremely dangerous and need to be monitored continuously having potential which may lead to death. There are different pollutants present in the air which may need to monitor for better living. There are some areas in the city which are highly polluted because of population, vehicles, industries etc. The polluted areas are very dangerous to human health which needs continuous monitoring. The air pollution is the introduction of particulates, biological molecules and other harmful materials into Earth's atmosphere which causes diseases to humans and damage to other living organisms such as animals and food crops or the natural or built environment. Air pollution can be caused due to various human activities such as industries, automobiles and burning of fossil fuels like wood, coal in thermal power plant project or naturally like wildfires. Though there is increase in the development of technology and human race but we have been failed to take care about the surroundings in which we live in. Thus environmental pollution reducing the quality of the air in the place people live into. One such example is Motor Vehicles. Due to motor vehicles aproximately 25% of the hazardous gases has released into air. Outdoor environment pollution levels are the key concern, but the quality of air inside the vehicle plays major part .The air pollution from vehicles in urban areas, particularly in big cities, has become a serious problem. With the increase in the number of vehicles due to urbanization, air pollution has increased rapidly in the past few years. The primary pollutants emitted from these automobiles are carbon monoxide, oxides of nitrogen and unburned hydrocarbons. CO is considered to be the most dangerous among all these. The health risks of air pollution are extremely seriousleading to various diseases such as cancer, asthma, Cardiovascular Disease, diabetes, bronchitis and also putting the elderly and the kids at a higher risk. As a result various measures are taken to reduce the Vehicular pollution. The factors that contribute to vehicular pollution are poor fuel quality, old vehicles, inadequate maintenance, old automotive technologies and traffic management. Thus vehicles that are more fuel efficient and those that produce fewer emissions are some of the means by which we can reduce transport related air pollution. Emission from vehicles cannot be completely avoided but, it definitely can be controlled.

2. RELATED WORKS

Vishal mahuli, Vishwanath Kulkarni, AnvithaRao, Adarsh Raj (2016)[1] research paper mainly concern with the prediction of air quality with machine learning. The air consists of different biological molecules, or other harmful materials which are harmful for the mankind.

The six “criteria pollutants” are ground level ozone(O3), fine particulate matter ($PM_{2:5}$), carbon monoxide (CO), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), and lead, among which ground level O_3, $PM_{2.5}$ and NO_2 (main component of NO_x) are the most widespread health threats.

The Air Quality Health Index (AQHI) is a tool designed in Canada to understand the air quality impact on human health. Basically this AQHI gives an index or rating scale range from 1 to 10 based on the health risk associated with local air quality. The most current air quality forecasting technique uses straightforward approaches like boxmodels, Gaussian models and linear statistical models. The advantages of those models are easy to implement and it allows for the rapid calculation of forecasts. However the main disadvantages of those techniques that they are not able to describe the interactions and non-linear relationship that control the transport and behavior of pollutants in our atmosphere

[5]. For this changes the Machine Learning methods are introduced which originating from the artificial intelligence in air quality forecasting and other atmospheric problems. [6] Several neural network (NN) models have already been used for air quality forecast such as forecasting hourly averages [7] and daily maximum [8]. Although NN have many advantages over traditional statistical methods for air quality forecasting. But these NN-based models still need to be improved to get the precise result as effectively as possible [9]. There are number of difficulties associated with NN hamper for their effectiveness in air quality forecasting. These difficulties are computational expense, multiple local minima during optimization, over-fitting to noise in the data etc. Therefore there are no general rules to determine the optimal size of network and learning parameters so according to the performance when the size of the network will increase then there will be a chance to get the precise result but the problem is the complexity will increase over the network size.

Another key consideration of forecast models is their updatability when the forecasting gives in real time. Normally there are two ways for model updating one is "batch learning" and the next one is "online learning". Whenever the new data are received then "batch learning" uses the past data together with the new data and performs a retraining process of the model and the other hand "online learning" uses only the new data to update for the model. The main disadvantage of "Batch learning" that it is considered as computationally expensive in real-time forecasting as the process means repeatedly altering a representative set of parameters calibrated over a long historical record. Linear models are generally easy to update online [10] and even with "batch learning" the result will come fast and easy to implement. As for non-linear methods the "online learning" is difficult because it uses many formulations such as the non-linear kernel method. So for the daily update or short time updates using "batch learning" is too expensive.

The research goal of this paper is to develop a non-linear updatable model works with real-time air quality forecasting using updatable linear regression models.

1.2.1 Neural network (NN) methods were developed from investigations into human brain function and these are adaptive systems that change over time as they learn [12].

In the figure 1 x_i is the inputs, h_j is the hidden neurons and y_k is the output. According to the research the neural network have the ability for good forecasting. [13] used NN methods to simulate the observed global (and hemispheric) annual mean surface air temperature variations during 1874-1993 using anthropogenic and natural forcing mechanisms as predictors.

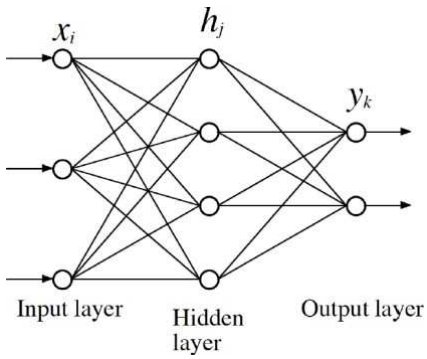


Figure 1: The general structure of a MLP NN model [11]

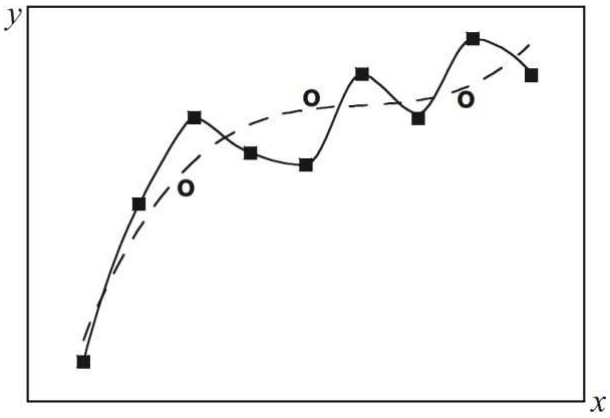


Figure 2: A diagram illustrating the problem of over-ftting [12]

NN having many benefits over the statistical approach but the main problem is over-fitting. According to the Figure 2 the solid square line is based on the training dataset which is quite good but the hollow circle line has over-fitted with respect to the training data so it leads to poor prediction.

1.2.2 Extreme Learning Machine (ELM) [14] [15] The ELM algorithm has the same architecture as a single-hidden layer feed-forward neural network (SLFN) but the advantage is that it is generally fast to train. The ELM works by randomly chooses the weights leading to the Machine Learning Techniques hidden nodes or neurons (HN) and analytically determines the weights at the output layer by solving a linear least squares problem. The only parameter that needs to be tuned in the ELM is the number of HN [16].

1.2.3 Air Quality Forecasting Models

An air quality model is a numerical tool used to describe the relationship among emissions, meteorology, atmospheric concentration, deposition and other factors. It can provide a complete deterministic description of the air quality problem [17]. The most commonly used air quality models are discussed below.

1.2.3.1 Dispersion Models normally use mathematical formulations to evaluate the atmospheric situation after pollutants were emitted by a source. It requires meteorological data, emissions data, and details in order to give facilities in stack height, gas exit velocity, etc. Some of the more complex models even require topographic information, chemical characteristics individually and land use data. The output prediction has concerned at selected locations. There are different types of dispersion models with specific requirement and special scales. The most commonly used dispersion models are the box model, Gaussian plume model, Lagrangian model, Eulerian model and Gaussian pu model [18].

1.2.3.2 Photochemical Models have become widely used in air pollution control strategies. Photochemical models identify the changes of pollutant concentrations in the environment using a set of mathematical equations characterizing the chemical and physical processes in atmosphere. These models are applied at different scaling factors from local, regional, national, and global [19].

1.2.3.3 Regression Models

Both linear regression and non-linear regression models have been employed for air quality forecasting. The general purpose of a linear regression model is to learn about the linear relationship between several independent variables (predictors) and a dependent variable (predictand). [20] Built a simple linear regression model for forecasting the daily peak O_3 concentration in Houston.

1.2.3.4 Neural Network Models

Although many approaches such as box models, Gaussian plume models, persistence and regression models are commonly applied to characterize and forecast air pollutants concentration, they are relatively straightforward with signicant simplications[20].

A. R. Al-Ali, Imran Zualkernan, and Fadi Aloul (2010) [2] discussed online GPRS-Sensors Array for air pollution monitoring. The author has been designed, implemented, and tested the proposed system. The proposed system consists of a Mobile Data-Acquisition Unit (Mobile-DAQ) and a fixed Internet-Enabled Pollution Monitoring Server (Pollution-Server). The Mobile-DAQ unit integrates a single-chip microcontroller, air pollution sensors array, a General Packet Radio Service Modem (GPRS-Modem), and a Global Positioning System Module (GPS-Module). The Pollution-Server is a high-end personal computer application server with Internet connectivity. The Mobile-DAQ unit gathers air pollutants levels (CO, NO2, and SO2), and packs them in a frame with the GPS physical location, time, and date. The frame is subsequently uploaded to the GPRS-Modem and transmitted to the Pollution-Server via the public mobile network. A database server is attached to the Pollution-Server for storing the pollutants level for further usage by various clients such as environment protection agencies, vehicles registration authorities, and tourist and insurance companies. The Pollution-Server is also interfaced to Google Maps to display real-time pollutants levels and locations in large metropolitan areas. The system was successfully tested in the city of Sharjah, UAE. The system reports real-time pollutants level and their location on a 24X7 .

Most of the air pollution and quality monitoring systems are based on sensors that report the pollutants levels to a server via wired modem, router, or short-range wireless access points. In this paper, the author has proposed a system that integrates a single-chip microcontroller, several air pollution sensors (CO, NO2, SO2), GPRS-Modem, and a general positioning systems (GPSs) module. The integrated unit is a mobile and a wireless data acquisition unit that utilizes the wireless mobile public networks. The unit can be placed on the top of any moving device such as a public transportation vehicle. While the vehicle is on the move, the microcontroller generates a frame consisting of the acquired air pollutant level from the sensors array and the physical location that is reported from the attached GPS module. The pollutants frame is then uploaded to the General Packet Radio Service Modem (GPRS-Modem) and transmitted to the Pollution-Server via the public mobile network. A database server is attached to the Pollution-Server for storing the pollutants level for further usage by interested clients such as environment production agencies, vehicles regeneration authorities, tourist and insurance companies. The Polution-Server is interfaced to Google maps to display real-time pollutants levels and their locations in large metropolitan area such as Sharjah City, UAE.

The paper specifies the system functional and nonfunctional requirements, system hardware, software architecture alogn with the results and implementation.

Shwetal Raipure (2014)[3] has implemented a model based on the wireless sensor network (WSN) technique. The system investigates the pollution of different areas in metropolitan cities. The author has designed the architecture of WSN based pollution monitoring system which collects pollution range from different areas and send the collected information to Server. Wireless sensors used to calculate the percentage of harmful gases present in the air which is useful to avoid different health related issues in that particular area by using Data Mining. Real time results have shown the good performance of the proposed calculation scheme compared to other traditional scheme. By calculating polluted air in different areas using wireless sensors can be useful to calculate percentage of harmful gases and ultimately reduce the pollution in air. The proposed system not only measures the pollutant level, temperature, humidity but also we can forecast possibility of future pollution range by using data mining algorithm to the database.

The author used various sensors to measure the percentage of pollutants present in the particular areas of the city. The

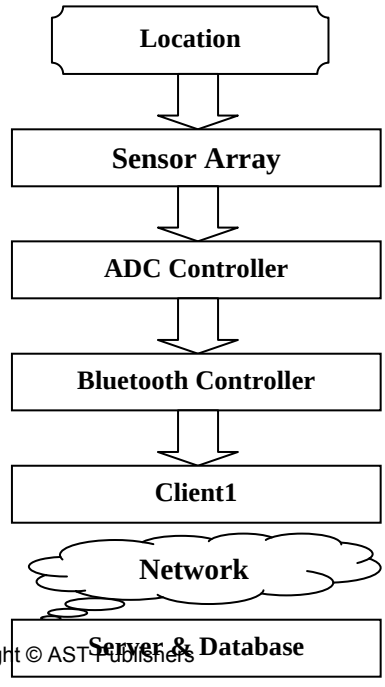
proposed model can also detect temperature and humidity present in the air using sensors. By using Bluetooth controller the collected data will send to server. Then it apply ID3 data mining algorithm which is useful for calculation of the percentage of pollutants in the air as well as temperature and humidity. With the help of data mining algorithm, the model will give future predictions to the particular area in the city and can also provide alarm to highly polluted area.

3.1 PROPOSED MODEL OUTLINE

1. Develop architecture to define nodes and their interaction.
2. Collect air pollution readings from different region of interest.
3. Collaboration among thousands of nodes to collect readings and transmit them to a gateway, which minimizes duplicates and invalid values.
4. Use of appropriate data aggregation to reduce the power consumption during transmission of large amount of data between the thousands of nodes.
5. Visualization of collected data from the WSN using statistical methods such as tables and line graphs.
6. Provision of an index to categorize the various levels of air pollution, which represent the seriousness of air pollution.
7. Generation of reports as well as real-time notifications during serious states of air pollution for use by appropriate authorities.

The Air Monitoring Unit in Mauritius makes use of bulky instruments which lacks resources. This reduces the flexibility of the system and makes it difficult for proper controlling and monitoring. This system will try to enhance this situation by being more flexible and timely. It will provide accurate data with indexing capabilities.

3.2 FLOW OF THE SYSTEM



Prof. Vishal V. Pande, Rupesh A. Kale, Rupali S. Shirke, Jigar V. Chitroda, Aakash P. Panchal (2015)[4] has developed a compact system to detect the pollutants in the vehicle which could be assembled in the vehicle itself. Tremendous innovations have been made in the technology and manufacturing of cars as well as in the pollution control department but still nothing significant achieved of it. This idea employs an MQ7 sensor which is economical and capable of detecting Carbon Monoxide gas emitted from the vehicle. An initial warning is given to the driver regarding the amount of CO gas with the help of LCD display and later the same information is transferred to the Pollution Control Board in case of negligence. This is done with the help of GSM system incorporated in the vehicle.

The AVR Microcontroller is used to transfer the information to the GSM system from the MQ7 sensor. The aim of this paper is to develop a mobile PUC checking system.

This paper comes to picture as it focuses on the idea of eradicating the existing PUC system totally and the above discussed are some of the research papers relating to the area of interest of our paper.

4.1 MQ-7 SENSOR

The Sensitive Material of MQ-7 gas sensor is SnO_2 which with lower conductivity in clean air. It make detection by method of cycle high and low temperature, and detect CO when low temperature (heated by 1.5V). The sensor's conductivity is higher along with the gas concentration rising. When high temperature (heated by 5.0V), it cleans the other gases adsorbed under low temperature. The sensor could be used to detect different gases contains CO, it is with low cost and suitable for different application. Good sensitivity to Combustible gas in wide range. This Sensor detects the presence of Carbon Monoxide at concentrations from 20 to 20,000 ppm. The sensor can operate at temperature from -10 to 50 C and consumes less than 150 mA at 5V.

4.2 AVR MICROCONTROLLER

The AVR Microcontrollers are low-power CMOS 8-bit controller based on the RISC architecture. The AVR core combines a rich instruction set with general purpose working registers. All the registers are directly connected to the Arithmetic Logic Unit (ALU), allowing two independent registers to be accessed in one single instruction executed in one clock cycle. The resulting architecture is more code efficient while achieving throughputs up to ten times faster than conventional CISC microcontrollers. The AVR is a modified Harvard architecture 8-bit RISC single chip microcontroller which was developed by Atmel in 1996. It was one of the first microcontroller families to use on-chip flash memory for program storage. In this Project we are using **Atmel's AT90S8535 Microcontroller.**

Conclusion and future scope

Future work for this project may include involving vast IoT system having efficient sensors for detecting other toxic gases accurately. The received sensor data can be further analysis to calculate the AQI (Air Quality Index).

SUMMARY

This survey consists of two different papers with same goal. The goal is to predict the Air Pollution; one author provides the result through Machine Learning in the other hand another author takes the sensor data (IoT) which containing pollution parameters and makes an algorithm to prevent the pollution emission gases by stopping the engine of the vehicles. So any researcher can take the idea from both the authors and merge those techniques to build up a mechanized tool which can able to find out the pollution causes sources and also prevent it.

3. REFERENCES

[1] Vishal mahuli, Vishwanath Kulkarni, AnvithaRao, Adarsh Raj Student VII SEM, B.E, Computer Science. Engg., MSRIT, Bangalore, IndiaBrown, IJRCCT, Vol 5, Issue- 7, July – 2016

[2] A Mobile GPRS-Sensors Array for Air Pollution Monitoring A. R. Al-Ali, Member, IEEE, Imran Zualkernan, and Fadi Aloul, Senior Member, IEEE, VOL. 10, NO. 10, OCTOBER 2010.

[3] Calculating Pollution in Metropolitan Cities using Wireless Sensor Network Shwetal Raipure, Volume 2, Issue 12, December 2014 International Journal of Advance Research in Computer Science and Management Studies.

[4] Online Vehicle Pollutants Monitoring System using GSM, Prof. Vishal V. Pande, Rupesh A. Kale, Rupali S. Shirke, Jigar V. Chitroda, Aakash P. Panchal, Vol. 4, Issue 4, April 2015.

[5] L. D., Hua, H., and Gao, C. 2003. A widget framework for augmented interaction in SCAPE.

[6] Y.T. Yu, M.F. Lau, "A comparison of MC/DC, MUMCUT and several other coverage criteria for logical decisions", Journal of Systems and Software, 2005, in press.

[7] Spector, A. Z. 1989. Achieving application requirements. In Distributed Systems, S. Mullender

[8] Luecken, D. J., Hutzell, W. T., and Gipson, G. L. (2006). Development and analysis of air quality modeling simulations for hazardous air pollutants. Atmospheric Environment,40(26):5087-5096.

[9] Comrie, A. (1997). Comparing neural networks and regression models for ozone forecasting. Journal of Air and Waste Management, 47:653-663.

[10] Kolehmainen, M., Martikainen, H., and Ruuskanen, J. (2001). Neural networks and periodic components used in air quality forecasting. Atmospheric Environment, 35(5):815-825.

[11] Perez, P., Trier, A., and Reyes, J. (2000). Prediction of PM2.5 concentrations several hours in advance using neural networks in Santiago, Chile. Atmospheric Environment, 34:1189-1196.

[12] Wang, W., Xu, Z., and Lu, J. W. (2003). Three improved neural network models for air quality forecasting. Engineering Computations, 20(2):192-210.

[13] Wilson, L. J. and Vall_ee, M. (2002). The Canadian updateable model output statistics (UMOS) system: design and development tests. Weather Forecast, 17:206-222.

[14] Hsieh, W. W. (2009). Machine Learning Methods in the Environmental Sciences: Neu-ral Networks and Kernels. Cambridge University Press.

[15] Hsieh, W. W. and Tang, B. (1998). Applying neural network models to prediction and data analysis in meteorology and oceanography. Bulletin of the American Meteorological Society, 79:1855{1870.

[16] Walter, A., Denhard, M., and Schonwiese, C.-D. (1998). Simulation of global and hemispheric temperature variations and signal detection studies using neural networks. Meteorologische Zeitschrift, N.F.7:171-180.

[17] Huang, G. B., Chen, L., and Siew, C. K. (2006a). Universal approximation using incremental constructive feedforward networks with random hidden nodes. IEEE Transactions on Neural Networks, 17(4):879-892.

[18] Schmidt, W. F., Kraaijveld, M. A., and Duin, R. P. W. (1992). Feed forward neural networks with random weights. In 11th IAPR International Conference on Pattern Recognition, Proceedings, Vol Ii: Conference B: Pattern Recognition Methodology and Systems, pages 1-4.

[19] Lima, A. R., Cannon, A. J., and Hsieh, W. W. (2015). Nonlinear regression in environmental sciences using extreme learning machines: A comparative evaluation. Environmental Modelling and Software, 73:175-188.

[20] Nguyen, D. (2014). A brief review of air quality models and their applications. Open Journal of Atmospheric and Climate Change, 1(2):60-80.

[21] Holmes, N. S. and Morawska, L. (2006). A review of dispersion modeling and its application to the dispersion of particles: An overview of different dispersion models available. Atmospheric Environment, 40(30):5902-5928.

[22] Nguyen, D. (2014). A brief review of air quality models and their applications. Open Journal of Atmospheric and Climate Change, 1(2):60-80.

[23] Prybutok, V. R., Yi, J., and Mitchell, D. (2000). Comparison of neural network model with ARIMA and regression models for prediction of Houston's daily

maximum ozone concentrations. European Journal of Operational Research, 122:31-40.

[24] Luecken, D. J., Hutzell, W. T., and Gipson, G. L. (2006). Development and analysis of air quality modeling simulations for hazardous air pollutants. Atmospheric Environment, 40(26):5087-5096.

[25] Vishal mahuli et al, IJRCCT, Vol 5, Issue- 7, July – 2016 “Effective Prediction And Prevention Of Air Pollution Caused Due To Automobiles Using IOT And Data Analytics Techniques” ISSN (Online) 2278- 5841

Predicting the Authenticity of Banknotes Using Supervised Learning

Priyam Guha
Computer Science
Institute of Engineering and Management
Kolkata, India
priyamguha@gmail.com

Abhishek Mukherjee
Computer Science
Institute of Engineering and Management
Kolkata, India
abhishek98mukherjee@gmail.com

Abhishek Verma
Computer Science
Institute of Engineering and Management
Kolkata, India
abhishek089verma@gmail.com

***Abstract*—This research paper deals with using supervised machine learning algorithms to detect authenticity of bank notes . In this research we were successful in achieving very high accuracy (of the order of 99%) by applying some data preprocessing tricks and then running the processed data on supervised learning algorithms like SVM , Decision Trees , Logistic Regression , KNN. We then proceed to analyze the misclassified points . We examine the confusion matrix to find out which algorithms had more number of false positives and which algorithm had more number of False negatives.[1] This research paper deals with using supervised machine learning algorithms to detect authenticity of bank notes . In this research we were successful in achieving very high accuracy (of the order of 99%) by applying some data preprocessing tricks and then running the processed data on supervised learning algorithms like SVM , Decision Trees , Logistic Regression , KNN. We then proceed to analyze the misclassified points . We examine the confusion matrix to find out which algorithms had more number of false positives and which algorithm had more number of False negatives.[1] .**

***Index Terms*—Supervised Learning, Support Vector Machine, Logistic Regression, K-Nearest Neighbors, Decision Trees**

I. Introduction

Monetary transactions exists from long back, from the times of ancient human civilizations. There has been various methods of monetary transactions. In the modern world two major forms of monetary transactions are cash transactions and electronic transactions. Till the previous decade cash transactions were the only means above all others, but there has been a decrease in the cash transaction due to the recent increase of electronic transactions. However, cash transactions are still very important part of the global market. Banknotes are used to carry out financial works across the entire nation forming the backbone for our economical growth. The global market has faced the entry of fake banknotes into the financial market in a huge rate in the recent times, hence for the smooth running of transactions and preventing any fraud transactions from being successful.[2] Fake notes are copies of actual notes manufactured through various technologically advanced copying machines. The fake notes are manufactured in all denominations which droops down the economical situation of a country to a very low level and the difficulty of their identification also increases. The recent advancement in the copying and scanning methods have led the tricksters to create the create fake notes in large numbers making this a very easy task for them. It becomes very difficult for a person to differentiate between a genuine and a fake note with the naked eye because of the close resemblance in features of the actual ones in comparison to the fake ones. Hence, there is an important need for an efficiently implemented accurate system in banks and ATM's for the classification of genuine notes from the fake ones.

In the recent years, the use of soft computing techniques have solved problems that were quite difficult to solve using conventional mathematical methods.[5] These techniques have better efficiency for classifying different unauthentic note being introduced in the financial market.For a practical situation, Consider a person wants to deposit money in the bank using bank currency notes. The notes that are to be deposited are given to a human being(a person of the bank) to check for their authenticity. As the fake notes are prepared with precision, it is difficult to differentiate them from genuine ones. An identification system must be installed to detect the legitimacy of the bank note.This paper evaluates few supervised machine learning algorithms, namely, Logistic Regression, KNN, decision trees and SVM for classification of genuine and fake notes. A comparison study is also presented where the proficiency of these well known algorithms are measured in the basis of accuracy, sensitivity, and specificity.

The data set used to train these algorithms was collected by extracting features from banknote images. The dataset also classifies all the samples into a particular class i.e. genuine or forged(from the Class attribute).

II. Dataset and Preprocessing

In the following study banknotes are being checked for their authenticity using a set of parameters obtained from a dataset. Data were extracted from images that were taken from genuine and forged banknote-like specimens. For digitization, an industrial camera usually used for print inspection was used. The final images have 400 x 400 pixels. Due to the object lens and distance to the investigated object gray-scale pictures with a resolution of about 660 dpi were gained. Wavelet Transform tool were used to extract features from images.

Data preprocessing is one of the most important aspect of machine learning . Data preprocessing is a technique that is used to convert the raw data into a clean data set. Data preprocessing also helps in achieving much better results with the same algorithms. Common techniques of data preprocessing includes :

Ignoring the missing record , Filling the missing values manually , Filing using computed values, Binnig method(Sorting of data is performed concerning the values of the neighborhood) and sometimes using correlation values to combine two columns in various ways and selecting the best features possible.

The dataset on which we are working has mainly five features namely VAR , SKEW , CURTOSIS , ENTROPY and CLASS. The below picture denotes the features of the dataset we have . we see that mean of all the features is well within -3 to +3 , hence there is no need for Feature Scaling. Now we perform pair wise correlation to find out which features has the most effect on the Class.

By experimenting with various combinations of features we found out that , by creating a new feature by simply adding the features CURTOSIS and SKEW we get a decision boundary which clearly separates the authenticate notes from the fake ones. The below picture depicts the clear decision boundary obtained just by adding the new feature.

This clear decision boundary between the new feature (q)(Skewed curtosis) and the VAR feature helps the supervised algorithms to achieve a very high accuracy. This modification in data helps us to achieve more than 99% accuracy in some of the algorithms which without using this preprocessing was around 60% in best conditions .

TABLE I
INITIAL DATASET INFORMATION

	Values in the Dataset				
Values	***VAR***	***SKEW***	***CURTOSIS***	***ENTROPY***	***CLASS***
Count	1372.00	1372.00	1372.00	1372.00	1372.00
Mean	0.4337	1.9224	1.3973	-1.1916	0.4461
Std.	2.8428	5.8694	4.3100	2.1010	0.4971
Min.	-7.0421	-13.7731	-5.2861	-8.5482	0.0000
25%	-1.7730	-1.7082	-1.5749	-2.4134	0.0000
50%	0.4962	2.3197	0.6166	-5.8665	0.0000
75%	2.8215	6.8146	3.1792	0.3948	1.0000
Max.	6.8248	12.9516	17.9274	2.4495	1.0000

[a]

The supervised learning models have been implemented in jupyter in Anaconda in python 3. The dataset used is divided into two subsets i.e. ratio of 70:30. The bigger subset is used for training the models and the smaller subset is used to test whether the models can predict the genuineness of note or not. An extra attribute named skewed curtosis was added to the dataset. Adding this feature gave a better prediction to the previous results.

III. IMPLEMENTED ALGORITHMS

In the following section different machine learning algorithms have been tested to get predictions whether a note is genuine or not. The following analyses are a part of detailed comparative study among different algorithms with an aim to obtain the best algorithm for the prediction of genuineness of bank notes.

A. Logistic Regression

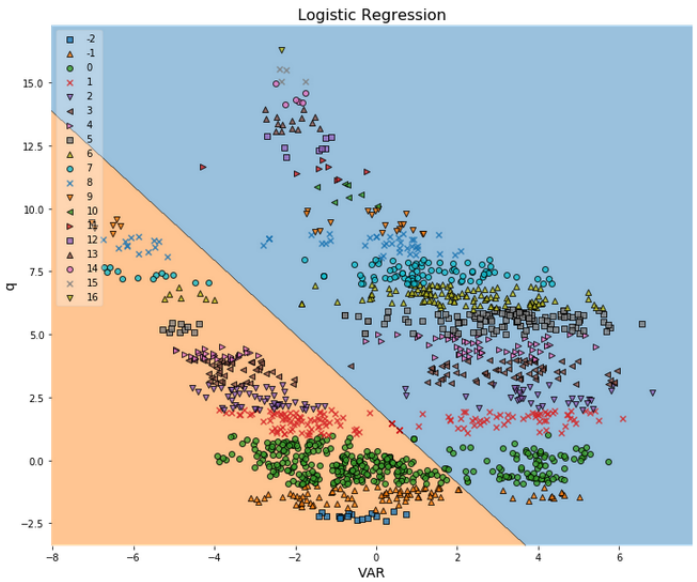
Logistic regression is one of the supervised learning algorithm which calculates the relationship between the class variable(the result which we want to predict) and other known attributes obtained from the datasets, by estimating probabilities using it's underlying logistic function.

The probabilistic relationship obtained between the different attributes present in the dataset and the class variable are transformed into binary values in order to make a prediction. the prediction is done with the help of the logistic function, also known as the sigmoid function.

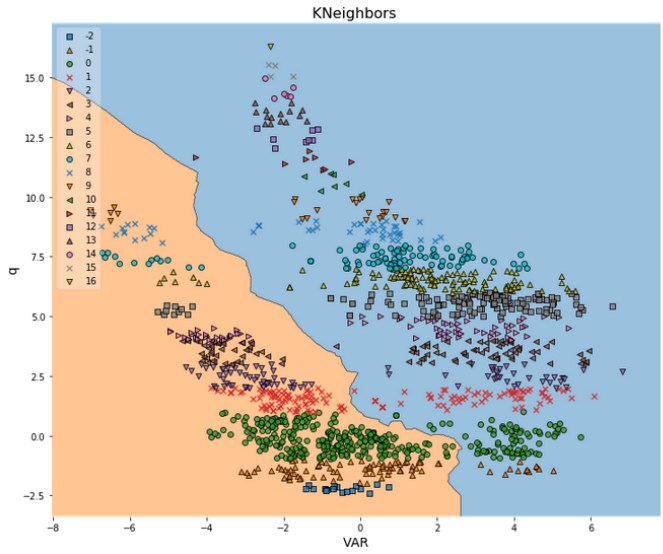
TABLE II
DATASET INFORMATION

Serial	**Final Dataset Informations**		
Number	***Attribute***	***Data Type***	***Description***
1	variance of Wavelet transformed image	continuous	Variance finds how each pixel varies from the neighbouring pixels and classifies them into different regions
2	skewness of Wavelet Transformed image	continuous	Skewness is the measure of the lack of symmetry
3	curtosis of Wavelet Transformed image	continuous	Curtosis is a measure of whether the data are heavy tailed or light tailed relative to a normal distribution
4	entropy of image	continuous	Image entropy is a quantity which is used to describe the amount of information which must be coded for, by a compression algorithm
5	skewed curtosis	continuous	an extra feature was added to the dataset, which was used for the prediction of real notes. Adding this feature gave a better prediction to the previous results
6	class	discrete integer	Class contains two values 0 representing genuine note and 1 representing fake note

[a]attribute 5 is a self added feature for better prediction result.



The Sigmoid-Function is an S-shaped curve that can take any real-valued number and map it into a value between the range of 0 and 1, but never exactly at those limits. This values between 0 and 1 will then be transformed into either 0 or 1 using a threshold classifier, which in turn classifies the class variable into one of the class and thus predicts the result. After implementing logistic regression on the test data the model which was generated, predicted the accuracy score of 98.06% on the class variable to classify the fake and genuine note. The result was further improved on implementation of more algorithms.



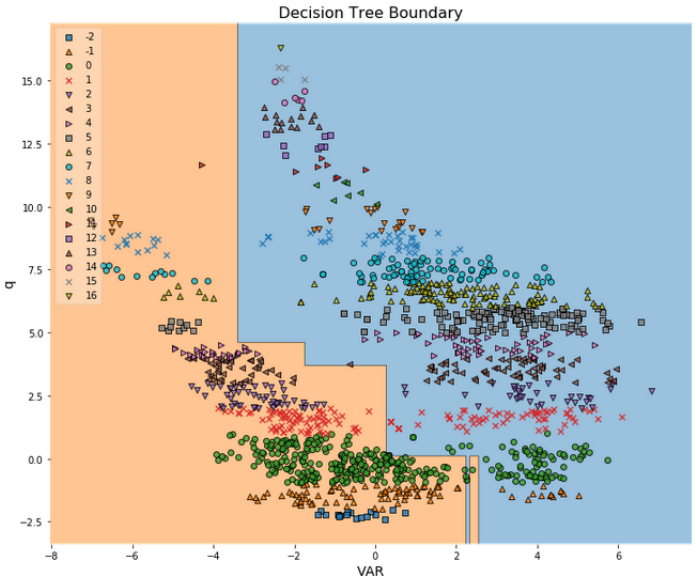
B. K- nearest neighbors

K nearest neighbors is a simple supervised learning algorithm that stores all available cases from the test data available and classifies new test data based on a grouping measure (e.g., distance functions). The test data falls under one of the available groups and is then classified, predicting the result. In kNN classification, the output is a class membership, which means that the predicted result falls under one of the categories of classification or class. An object is classified by a majority vote of its neighbors(using the grouping techniques), with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, which is small and a odd number). If k = 1, then the object is simply assigned to the class of that single nearest neighbor. Thus, the predicted result depends on the maximum number of data having similarities among the k-closest data members.

After the implementation of KNN algorithm the model which was generated predicted the an accuracy score of 99.03% with the implemented data. This result was further improved with the implementation of other algorithms on the test data.

C. Decision trees

Decision tree algorithm uses a decision tree (as a predictive model implemented on the training data to predict the result of the test data) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaf nodes) It is one of the predictive modeling approaches used in statistics, data mining and machine learning.In these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees.[4]



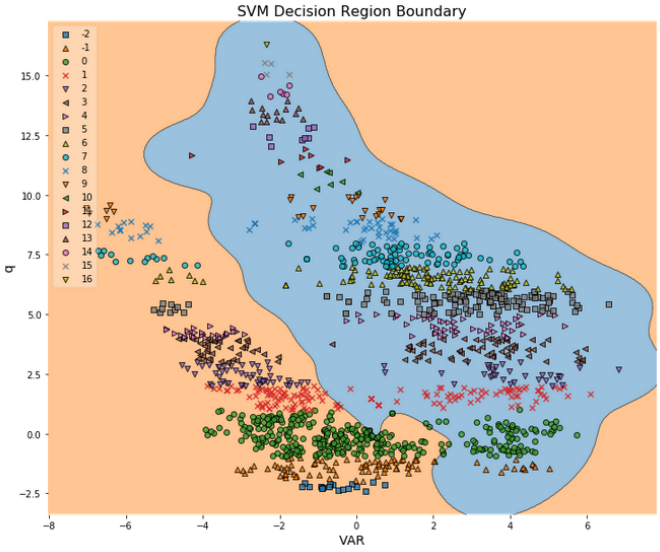
Decision Tree Classifier, iteratively divides the working area(plot) into sub part by identifying lines. (repetitively because there may be two distant regions of same class divided by other.) Thus, classifying the test data into different groups ultimately predicting the class in which the test data belongs.

By the use the classifier decision trees we got an accuracy of prediction of 99.51%. this accuracy is further increased in the next used algorithm. Thus the prediction result of the decision trees is greater than that of KNN and logistic regression for the differentiation of fake notes and real ones.

D. support vector machine

In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning algorithms which are used for the prediction of both classification and regression problems having their related test and training test data. When a dataset is chosen having a class variable, where each dataset is marked as belonging to one of the different categories present in the class variable, making it a non-probabilistic binary linear classifier, a very different algorithm or actually opposite to the logistic regression algorithm, although methods exist to use SVM in probabilistic classification as in plot scaling.[3]

A predictive model generated after the implementation of SVM on a training dataset are modeled as different points in space, mapped so that the example of separate categories or classes are visibly divided by clear gap and which are as widely separated as possible for a better predictive result. New test data are then mapped onto the previously generated model and then they are predicted to which group the test data falls under. SVMs can efficiently perform linear classifications and also non-linear classification using kernel trick (implicitly mapping their inputs into high-dimensional feature spaces of different groups).



The model generated after the implementation of SVM on the training data predicted the accuracy score score of 99.64% on the test data, thus giving the best prediction result among the other implemented Supervised Learning algorithms. Hence, it can end our search for a better prediction algorithm in the case of distinguishing of fake and genuine bank notes.

IV. Conclusions

In the following paper we have used four different supervised learning algorithms to predict the genuineness of bank notes. The dataset had 5 attributes from which a new attribute skewed curtosis was generated using the sum of two attributes skewness and curtosis. After implementing different equations the sum proved to give the best result.The dataset didn't require to be scaled because of the near equal values of different attributes. The result table shows the different values of predicted results. When precision is high the number of false positives are high and when recall is high the number of false negatives are high. Hence here for the work of authenticating the bank notes we require less number of false negatives to decrease the nuber of fake notes being rendered as true ones. Thus in that case we need high precision value which was obtained in SVM, also The accuracy score of SVM was highest and was equal to 99.64% giving the best prediction, which is the best prediction result achieved regarding this dataset and hence can be soon put to effect.

TABLE III
RESULTS

Serial Number	Informations				
	Algorithm	***Accuracy***	***Precision***	***Recall***	***Remarks***
1	Logistic regression	98.06%	0.9808	0.993	the accuracy obtained from the generated model was least in comparison to the other implemented algorithms
2	K-Nearest Neighbors	99.03%	0.991	1	Using KNN the accuracy score obtained from the generated model increased in comparison to Logistic Regression
3	Decision Trees	99.51%	0.993	1	The Decision tree classifier had a better prediction result in comparison to the previously used algorithms
4	SVM(Support Vector Machine)	99.64%	1.0	0.993	SVM gave the best prediction result in comparison to others

[a]

References

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on machine learning* (pp. 161–168).

Chaum, D. (1983). Blind signatures for untraceable payments. In *Advances in cryptology* (pp. 199–203).

Hsu, C.-W., & Lin, C.-J. (2002). A comparison of methods for multiclass support vector machines. *IEEE transactions on Neural Networks*, *13*(2), 415–425.

Michalski, R. S., Carbonell, J. G., & Mitchell, T. M. (2013). *Machine learning: An artificial intelligence approach*. Springer Science & Business Media.

Mitra, S., & Acharya, T. (2005). *Data mining: multimedia, soft computing, and bioinformatics*. John Wiley & Sons.

A SURVEY ON CLOUD-DENIAL OF SERVICE

Bibek Naha
Institute of Engineering & Management, Kolkata
Email: bibek.naha123@gmail.com

Siddhartha Banerjee and Sayanti Mondal
Institute of Engineering & Management, Kolkata
Email: {sbanjee.banerjee, sayantiju2014}@gmail.com

Abstract — Cloud Computing is one of the most nurtured as well as debated topic in today's world. Billions of data of various fields ranging from personal users to large business enterprises reside in Cloud. Therefore, availability of this huge amount of data and services is of immense importance. The DOS (Denial of Service) attack is a well-known threat to the availability of data in a smaller premise. Whenever, it's a Cloud environment this simple DOS attack takes the form of DDOS (Distributed Denial of Service) attack. This paper provides a generic insight into the various kinds of DOS as well as DDOS attacks. Moreover, a handful of countermeasures have also been depicted here. In a nutshell, it aims at raising an awareness by outlining a clear picture of the Cloud availability issues.
Our paper gives a comparative study of different techniques of detecting DOS.

Index Terms— DDOS attack, Cloud Computing, Impact of DDOS attack.

I. INTRODUCTION

Cloud computing is the use of various services, such as software development platforms, servers, storage and software, over the Internet, often referred to as the "cloud." In general, there are three cloud computing characteristics that are common among all cloud-computing vendors:
The back-end of the application (especially hardware) is completely managed by a cloud vendor. A user only pays for services used (memory, processing time and bandwidth, etc.). Services are scalable. Google first put the concept of "cloud computing" formally in 2006 which opened a door to the research and practice of cloud computing. It also the reliable, cheap and convenient services for the users, which will reduce the user's hardware resources, software licenses and system maintenance costs.

DDoS attack refers to that these attackers run out the target resource or network bandwidth with the help of puppet hosts, to prevent the target from providing services to legitimate users. The serious threat to cloud computer is due to DDOS which is difficult to track as the attack sources are distributed widely in the real network environment. The traditional DDoS attack detection technology focuses on the characteristics and information of the target being attacked, and predicts and detects attacks through changes of characteristics and information of the target being attacked. The characteristics of the cloud environment

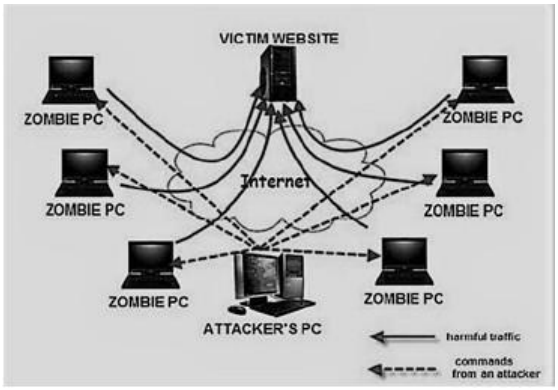


Fig 1: Defending against DDOS attacks [13]

Allow administrator to capture information of virtual machines and the corresponding network behaviours which just provides a clue for us to detect and prevent DDoS attacks.
Since the virtual machines deployed on the same node can share the same IP address, the traditional detection methods become difficult to identify the source of the transmitted attack traffic. In order to protect the cloud environment and prevent cloud resources from being consumed maliciously, this paper presents a comparison of different DDoS attack detection method based and the efficient technology used.

The rest of the paper is organized as follows, section 2 describes the Cloud Architecture, section 3 gives a brief overview of Denial of Service Attack, section 4 methods of DOS attack, section 5 tells us the impact of DOS attack and section 6 contains detection method of DOS attack.

II. CLOUD COMPUTING- A CONCEPT

Cloud computing is where computing resources are accessed from a virtual online cloud rather than a local desktop or organizational data Centre. Cloud

computing by mechanical definition refers to manipulating, accessing the application online which offers online data storage and also a combination of software and hardware based computing resource.

Accessing of cloud is done by following a model known as Deployment Model. The type of access to the cloud:-

Public cloud:- The Public Cloud allows systems and services to be easily accessible to the general public. Ex:- e-mail.

Private cloud:- The Private Cloud allows systems and services to be accessible within an organization.
Community cloud:- The Community Cloud allows systems and services to be accessible by group of organizations.

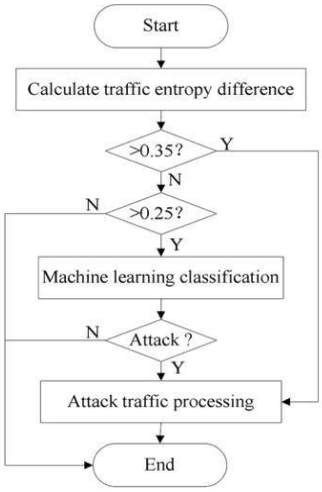


Fig 2: Truth table to show DOS attack [3]

Hybrid cloud:- The Hybrid Cloud is a mixture of public and private cloud. However, the critical activities are performed using private cloud while the non-critical activities are performed using public cloud.

Service Models are the reference models on which the cloud computing is based. The different types of Service models are:-
Infrastructure as a Service (IaaS):- Service provider offers capacity for rent basically hosted Data centres and Servers. Ex:- Rackspace.

Platform as a Service (PaaS):- Hosted application environment for developing and deploying cloud based applications. Ex:- Google's App Engine.

Software as a Service (SaaS):- In this case the application itself is provided by the service provider, typically via web browser. Ex:- Gmail

III. WHAT IS DENIAL OF SERVICE?

A DoS attack in simple language is scenario where a group of people are crowding the entry door of a shop, making it hard for original customers to enter and disrupting trade.
In computing, a denial-of-service attack (DoS attack) is a cyber-attack in which the perpetrator seeks to make a machine or network resource unavailable to its intended users by temporarily or indefinitely disrupting services of a host connected to the Internet. Denial of service is typically accomplished by ooding the targeted machine or resource with superuous requests in an attempt to overload systems and prevent some or all legitimate requests from being fulfilled.

IV. METHODS OF DOS ATTACK

4.1 VULNERABILITY-BASED ATTACK

This kind of attacks increases the aw in protocol design and defects the soft-ware. Under this kind of attack the service provided by the victim will be shut.

1. ICMP FLOOD

Smurf attack is a type of vulnerability based attack that occurs on the Internet services. There are services such as the Smurf Amplifier Registry that have provided the ability to the internet service providers to fight against Denial of Service attacks by identifying the networks with incorrect configuration and by filtering. Ping flood is a method that relies on sending a large number of ping packets to the victim. Ping of death is another method that is based on sending a malformed ping packet to the victim, as a result of which the system can crash.

2. SYN FLOOD

SYN flood is a result of TCP/SYN packets flooding sent by host, mostly with a fake address of the sender. Actually the sender never responds as his address is not real. The saturation of available connections takes place by the semi-open connections that the server can actually make, so that it cannot respond to legal requests even after the attack is over.

3. TEARDROP ATTACKS

In case of a Tear-drop attack the injured IP fragments are sent to the target machine with expanded, overlapping, payloads. As there is a bug in the TCP/IP fragmentation re-assembly code so this can result in crashing di erent operating systems.

4. LOW-RATE DENIAL-OF-SERVICE ATTACKS

The Low-rate DOS (LDOS) this type of attack actually exploits the TCPs slow-time-scale dynamics of re-transmission time-out (RTO) mechanisms so that it reduces TCPs output. Attacker can make the repeated entry of a TCP ow to a RTO state as the attacker can send the bursts at high-rate within short-duration, and this can be repeated periodically at slower

Re-transmission time-out time-scales. This results in reduced output of TCP.

4.2 FLOOD-BASED ATTACK

Flooding is a Denial of Service (DOS) attack that is designed to bring a network or service down by flooding it with large amounts of traffic. Flood attacks occur when a network or service becomes so weighed down with packets initiating in-complete connection requests that it can no longer process genuine connection requests. By flooding a server or host with connections that cannot be completed, the flood attack eventually fills the host memory buffer. Once this buffer is full no further connections can be made, and the result is a Denial of Service.

4.3 ZOMBIE ATTACK

A computer that has been implanted with a daemon that puts it under the control of a malicious hacker without the knowledge of the computer owner. Zombies are used by malicious hackers to launch DoS attacks. The hacker sends commands to the zombie through an open port. On command, the zombie computer sends an enormous amount of packets of useless information to a targeted Web site in order to clog the site's routers and keep legitimate users from gaining access to the site. The traffic sent to the Web site is confusing and therefore the computer receiving the data spends time and resources trying to understand the influx of data that has been transmitted by the zombies. Compared to programs such as viruses or worms that can eradicate or steal information, zombies are relatively benign as they temporarily cripple Web sites by flooding them with information and do not compromise the site's data. Zombies are also referred to as zombie ants.

4.4 REFLECTOR ATTACK

The reflector attack is, therefore, by its nature, more detrimental than using the zombie attack model alone because: It amplify the effect of the DDoS attack. Let us imagine that the attacker has only one zombie. By sending spoofed packets to different reflector, one zombie is already enough to attack the victim in a distributed way; It also degrades the services provided by the reflector. During the reflector attack, the reflector are loaded by the requests from the zombies, and this degrades the services provided by the reflector.

4.5 PEER-TO-PEER ATTACK

While peer-to- peer attacks are easy to identify with signatures, the large number were of IP addresses that need to be blocked means that this type of attack can overwhelm mitigation defences. Even if a mitigation device can keep blocking IP addresses, there are other problems to consider. For instance, there is a brief moment where the connection is opened on the server side before the signature itself comes through. Only once the connection is opened to the server can the identifying signature be sent and detected, and the connection torn down. Even tearing down connections takes server resources and can harm the server. This method of attack can be prevented by specifying in the peer-to-peer Protocol which ports are allowed or not. If port 80 is not allowed, the possibilities for attack on websites can be very limited [11].

4.6 WORM ATTACK

The worm attack is another form of automatic attack tool. To define, a worm is a piece of software that runs on a computer, and the computer is unwillingly having the worm running. The worm has the ability to duplicate itself, and has the duplicated copies infect other computers. Many worms that have been created are designed only to spread, and do not attempt to change the systems they pass through. However, even these "payload free" worms can cause major disruption by increasing network traffic and other unintended effects. A "payload" is code in the worm designed to do more than spread the worm it might delete files on a host, encrypt files in a crypto-viral extortion attack, or send documents via e-mail. A very common payload for worms is to install a backdoor in the infected computer to allow the creation of a "zombie" computer under control of the worm author [9][10][12].

Attack	Counter-measure Options	Example	Description
Network Level Device	Software patches, packet ltering	Ingress and Egress Filtering	Software upgrades can x known bugs and packet l-tering can prevent attack-ing tra c from entering a network
OS Level	SYN Cook-ies, drop backlog con-nections, shorten timeout time	SYN Cookies	Shortening the backlog time and dropping back-log connections will free up resources. SYN cook-ies proactively prevent at-tacks.
Application Level At-tacks	Intrusion Detection System	Guard Dog, other ven-dors	Software used to detect il-licit activity
Data Flood (Ampli-cation, Oscilla-tion, Simple Flooding)	Replication and Load Balancing	AkamiDigital Island pro-vide content distribution.	Extend the volume of con-tent under attack makes it more complicated and harder for attackers to identify services to attack and accomplish complete attacks.
Protocol Feature Attacks	Extend protocols to support security	ITEF stan-dard for itrace, DNSSEC.	Trace source destination packets by a means other than the IP address (blocks against IP address spoo ng). DNSSEC would provide authoriza-tion and authentication on DNS information.

Table 1: Different types of DOS attack [13]

V. IMPACT OF DDOS ATTACK

The impact of DoS attacks can vary from minor inconvenience to use of a web-site, to serious financial losses for companies that rely on their on-line availability to do business. DoS attacks generally occur basically in improper system design, insufficient resource. [4][5][6]

5.1 REVENUE LOSSES

Downtime affects bottom line of a system. Based on industry surveys, the average cost of downtime is 5, 600/minute, or over 300K/hour.

5.2 PRODUCTIVITY LOSS
When critical network systems are shut down, workforce's productivity comes to a halt.

5.3 REPUTATION DAMAGE
Your brand suffers if customers can't access your site or become casualties of a data breach.

5.4 THEFT
Attacks are becoming more advanced and now include stolen funds, customer data, and intellectual property.

VI. DDOS ATTACK DETECTION METHOD BASED ON TRAFFIC ENTROPY AND NAIVE BAYES

In order to protect the cloud environment and prevent cloud resources from being used maliciously, this paper discusses a DDoS attack detection method based on Traffic Entropy, Naive Bayes and Kolmogorov Complexity Method. By calculating the traffic entropy of the node of the virtual machine, combined with the machine learning classifier, the suspicious attack traffic can be identified and detected. Firstly, traffic entropy of virtual machines is

calculated, and the large traffic DDoS attack is identified by the entropy change. Then, the suspicious traffic is detected by machine learning to identify small ow DDoS attacks.

6.1 Traffic Entropy

According to characteristics of DDoS attacks, once multiple infected virtual machines of a cloud constitute a botnet to attack target victims, it will result in an increase in the number of packets targeted at the victim. We designate different packet to a destination address encountered in incoming packet, i.e. if the destination address is different, the packet is different. In a period of time, different destination addresses have different number of packets, i.e. the packet ratio is different and there is a certain range of changes. Once beyond the range, it is treated as a DDoS attack, so we can use the traffic entropy to measure packet changes for each destination address.

Classifiers	Accuracy	Sensitivity	Specificity
Naive Bayes	0.94	0.95	0.94
SVM	0.88	0.92	0.93
K-nearest	0.84	0.91	0.92

Table 2: Comparison between different detection methods [3]

6.2 Naive Bayesian Classifier

Merely considering the traffic entropy fluctuation in detecting small ow DDoS attacks, it is likely to confuse the normal traffic fluctuation with the flow fluctuation caused by the attack.

In order to identify small ow DDoS attacks in the cloud accurately, this paper proposes a machine learning detection algorithm based on naive Bayesian to detect the suspicious traffic. Traffic entropy difference in a certain period of time is used as a feature to identify the DDoS attacks with low ow entropy fluctuations. Naive Bias is a simple and efficient classification algorithm, which can compute the conditional probabilities efficiently.

6.3 Kolmogorov Complexity Method

The Kolmogorov Complexity, K(x), of a string of data measures the size of the smallest program capable of representing the given piece of data. It measures the degree of randomness for the given data. The length of the shortest program to generate a completely random string is equal to the size of the string itself. For all other cases, it is smaller than the size of the string and the program size becomes smaller as more regularity or pattern is discernible from the string. A side effect of this measure is its ability to represent the correlation between disparate pieces of data. This side effect is exploited to design an effective method for detecting DDoS attacks.

The DDoS attack detection algorithm makes use of a fundamental theorem of Kolmogorov Complexity [7] that states: for any two random strings X and Y,
$K(XY) <= K(X) + K(Y) + c$ [8] (1)

Where K(X) and K(Y) are the complexities of the respective strings, c is a constant and K(XY) is the joint complexity of the concatenation of the strings. Simply put, the joint Kolmogorov complexity of two strings is less than or equal to the sum of the complexities of the individual strings. The equivalence holds when the two strings X and Y are totally random i.e. they are completely unrelated to each other. Another effect of this relationship is that the joint complexity of the strings decreases as the correlation between the strings increases. Intuitively, if two strings are related, they share common characteristics and thus common patterns.

VII. CONCLUSION

Internet Services have become more important in which data is seamlessly ex-changed between client and server; therefore, Internet security is the essential component and DOS attack is the serious threat to the availability of the data. In this paper we have systematically analysed various DOS attack prevention techniques to deal with DOS attack under any circumstances and compared them in simulation environment, so we can find out which one is better to prevent DDOS attack when it happens.[2].

REFERENCE

[1] J. Cao, B. Yu, F.Dong, X Zhu and S. Xu, "Entropy -Based Denial of Service Attack Detection in Cloud Data Centre,"2014 Second International Conference on Advanced Cloud and Big Data, Huangshan, 2014, pp .201 - 207.

Doi : 10.1109/CBD.2014.34
URL : http://ieeexplore.ieee.org/stampstamp.jsptp=
arnumber=7176094 isnumber=7176054
[2] R.K. Sanodiya, "DoS attacks: A simulation study," 2017 International Conference on Energy, Communication, Data Analytics and Soft Comput-ing (ICECDS), Chennai, 2017, pp. 2553-2558. doi: 10.1109/ICECDS.2017.8389914
URL:http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=
arnumber=8389914isnumber=8389494
[3] W. Yang and D. Wei, "A distributed denial of service attack sources de-tection technology for cloud computing," 2017 4th International Confer-ence on Systems and Informatics (ICSAI),Hangzhou,2017,pp.660-664. doi: 10.1109/ICSAI.2017.8248371
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=
arnumber=8248371isnumber=8248252
[4] Monika Sachdev, Guruvinder Singh, Krishnan Kumar And Kuldip Singh.
[5] M.Robinson, J.Mirkovic,M.Schaider, S.Michel,and P.Reiher, Challenge and principle of the defense, Computer Journel of ACM SIGCOMM, vol.5,no.2,pp.148-152,2003.
[6] Damiano Bolzoni and Sandro Etalle. Boosting web intrusion detection system by inferring positive signature. In OTM Conference(2), pages 938-955,2008.
[7] A.B. Kulkarni, S.F. Bush, and S.C. Evans Detecting Distributed Denial-of-Service Attacks Using Kolmogorov Complexity Metrics, General Elec-tric Company, December 2001, 2001CRD176
[8] Rajkumar Buyya, Rodrigo N. Calheiros1, and Xiaorong Li. Autonomic Cloud Computing: Open Challenges and Architectural Elements, Cloud Computing and Distributed Systems (CLOUDS) Laboratory Department of Computing and Information Systems The University of Melbourne, Australia
[9] Sarika Agarwal, Saumya Agarwal, Bryon Gloden, DDOS Attack Simu-lation Monitoring, and Analysis
[10] Igino Corona, Giorgia Giancinto, Detection of Service-Side Web Attacks
[11] Ketki Arora, Krishan kumar, Monika Sachdev, Impact Analysis of DDOS Attacks, International Journal on Computer Science and Engineering(IJCSE).
[12] J.Howard, and T. Longsta ,a common language for computer security incident,[online].
[13] Mrs.S.Thilagavathi, and Dr.A.Saradha, Impact Analysis of Dos DDos Attacks. IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 16, Issue 6, Ver. IV (Nov Dec. 2014), PP 24-33 www.iosrjournals.org
[14] Shiva kumar , Ritika singal, priyadharshini , Mitigate the Impact of DOS Attack by Verifying Packet Structure, ECE Department LCET Katani, Kalan, India

Eradication Of Thalassemia By X-ray Photoelectronspectroscopy&DNA Spectral Analysis.

Annesha Nayak
Institute of Engineering &
Management
Kolkata,India
Email -id: annesha.nayak@gmail.com

Abstract:Chromosome no 11 & 16 of human embryo consist of the defective genetic sequence of alpha & beta thalassemia trait respectively.Here we want to eradicate the thalassemia by systematic method of analysing the defective genetic sequence of the chromosome no 11 & 16 if the conceiving couple are found to be carriers. This is further done amniocentesis by X-rayphotoelectronspectroscopy&D.N.A spectral analysis(that is done by decoding of the graph obtained from spectral analysis using computer algorithms, digital signal processing).

Keywords:

Thalassemiatrait,prenatal-test,amniocentesis,X-photoelectron spectroscopy,DNA Spectral analysis- which decodes the graph of defective genetic sequence fromthe power spectrum obtained from D.NA spectral analysis.

INTRODUCTION

Thalassemia is a genetic blood disorder. People with Thalassemia are not able to make enough haemoglobin which causes severe Anaemia. There are two types of Thalassemia traits: Alpha thalassemia trait and Beta thalassemia trait.

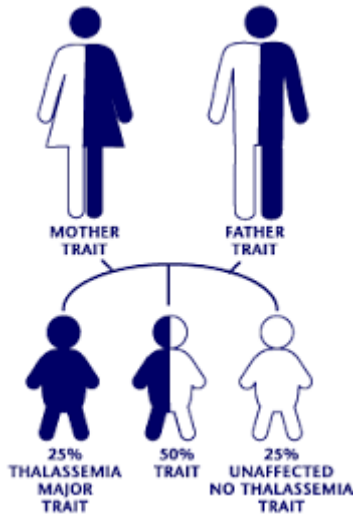


Fig 1:Inheritance pattern for thalassemia

ABOUT CARRIER :

I. ALPHA THALASSEMIAMINOR

Alpha thalassemia is responsible for deletion of D.N.A sequence of chromosome number 16.There are two different types of alpha thalassemia trait.The first type of alpha thalassemia trait has one alpha gene missing on each chromosome. This is called the ***trans*** form of alpha thalassemia trait.The second type of alpha thalassemia trait has two missing alpha genes on the same chromosome . This is called the ***cis*** form of alpha thalassemia trait.

II. ALPHA THALASSEMIA MAJOR

Alpha Thalassemia major occurs when both the parents are carriers. It is a form of thalassemia involving genes HBA1 & HBA2.Alpha -thalassemia is due to impaired production of alpha globin chains from 1,2,3 or all 4 of

the alpha globin genes ,leading to a relative excess of beta globin chains. The degree of impairment is based on which clinical phenotype is present (how many genes are affected).

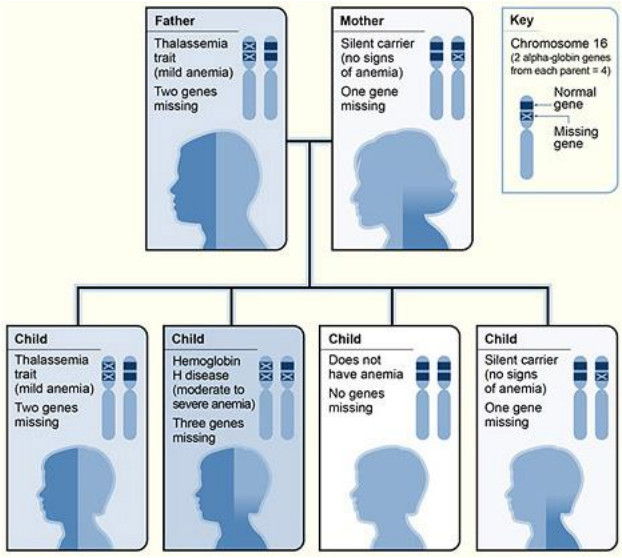


Fig 2:A typical inheritance pattern of Alpha Thalassemia Major

III. BETA THALASSEMIAMINOR

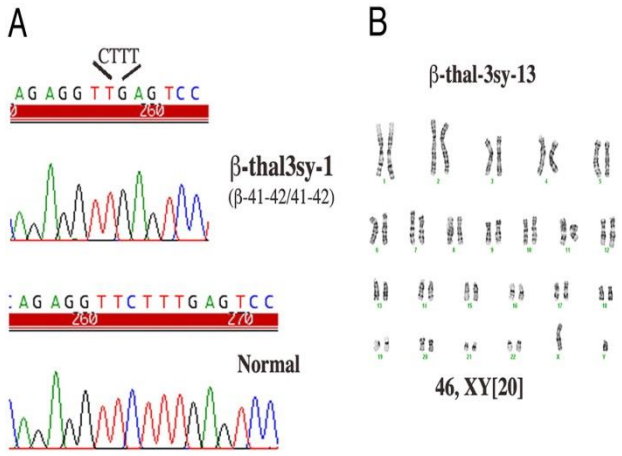


Fig 3 : Genetic sequence of chromosome number 11 having beta thalassemia trait& genetic sequence of normal chromosome 11.

Betathalassemia is responsible for mutation of D.N.A sequence of chromosome number 11. When someone has beta thalassemia ,there is a mutation in chromosome 11.Beta globin is made on chromosome 11(beta globin, along with alpha globin ,is one of the proteins that makes up haemoglobin).So, if one or both of the genes that tells chromosome 11 to produce beta globin is altered, less beta globin is made.

Normally , beta thalassemia trait doesnot cause any health problems.Beta thalassemia trait is also known as beta thalassemia minor.If one parent has beta thalassemia trait & the other parent has normal haemoglobin , there is a 50%(1 in 2) chance with each pregnancy of having a child with beta thalassemia trait.These are the possible outcomes with each pregnancy.

1)50% (1 in 2) chance of having a child with beta thalassemia trait.

2)50%(1 in 2) chance of having a child without trait.

IV.BETA THALASSEMIA MAJOR

Beta thalassemia major occurs when both the parents are carrier. It is a blood disorder that reduces the production of haemoglobin .Hemoglobin is the iron-containing protein in red blood cells that carries oxygen to cells throughout the body.In people with beta thalassemia, low levels of haemoglobin lead to a lack of oxygen in many parts of the body. Affected individuals also have a shortage of red blood cells which can cause pale skin, weakness, fatigue, more serious complications. People with increased beta thalassemia are at an increased risk of forming abnormal blood clots.

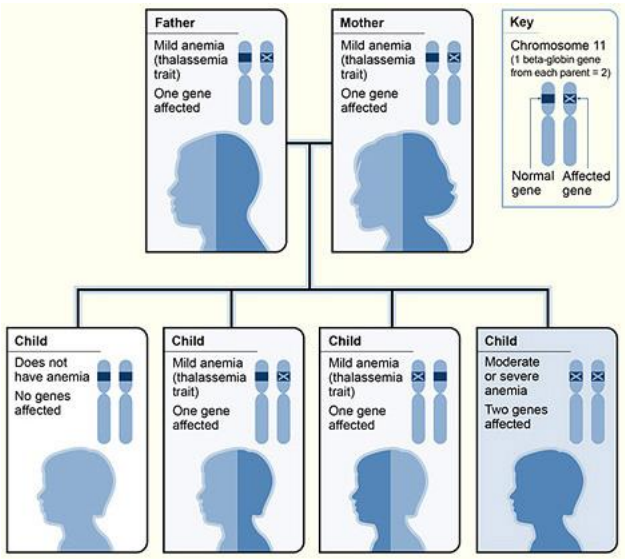


Fig 4: A typical inheritance pattern of Beta Thalassemia Major

IV. USES

Detection of Alpha / Beta thalassemia trait in humans:

Blood test - electrophoresis technique is used to determine the type of trait in humans that is alpha or beta thalassemia trait respectively. This blood test is done by couples before pregnancy.

Prenatal Screening is done for testing of any diseases or conditions in a foetus or embryo before it is born.This can be done by different processes but the process used here is amniocentesis.Amniocentesis is a medical procedure used in prenatal diagnosis of chromosomal abnormalities.This uses a small amount of amniotic fluid, which contains foetal tissues, is sampled from the amniotic sac surrounding a developing foetus and the foetal DNA is examined for genetic abnormalities. The most common reason to have an "amnio" is to determine whether a baby has certain genetic disorders or a chromosomal abnormality.Amniocentesis is usually done when a woman is between 14 and 16 weeksof pregnancy.The micrograph of the tissue is obtained from the embryo using X-ray photoelectron spectroscopy is then made to undergo D.N.A spectral analysis.

Quantitative characterization of D.N.A films using X-ray photoelectron spectroscopy:

X-ray photoelectron spectroscopy (XPS) is a surface-sensitive quantitative spectroscopic technique that measures the elemental composition at the parts per thousand range, chemical state and electronic state of the elements that exist within a material. XPS spectra are obtained by irradiating a material with a beam of X-rays while simultaneously measuring the kinetic energy and number of electrons that escape from the top 0 to 10 nm of the material being analyzed.

In D.N.A X-Ray photoelectronspectroscopy,the ray which is used is X-Ray.This ray uses the self-assembled films of thiolated 25 single – stranded D.N.A on gold as a model system for quantitative characterization of D.N.A films .We evaluate the applicability of an uniform & homogenous over layer – substrate model for data analysis, examine model parameters used to describe D.N.A films(e.g. density ,electron attenuation,length) validate the results. The model is used to obtain quantitative composition & coverage information as a function of immobilazationtime. We find that when the electron attenuation effects are properly included in the XPS data analysis excellent agreement is obtained with Fourier Transform infrared (FTIR)measurements for relative values of the DNA coverage, and the calculated absolute coverage is consistent with a previous radiolabelling study based on the effectiveness of the analysis procedure for model 25(ss)DNA films, it should be generally valid for direct quantitative comparison of DNA films prepared under widely varying conditions.

A. Application of D.N.A spectral analysis:

Spectral Analysis of D.N.A sequences reveals genome's periodicities from where a D.N.A power spectrum of the D.N.A sequences is plotted by analysing the hidden features of a D.N.A sequence by parametric method. It has been observed that in most cases data sequence vary with time but in few situations datamay vary with location points in space. Though in DNA sequences the variation is in position ofnucleotide bases, it is treated as a time-series signal. From point of view of statistics such sequencesare termed as Categorical time series. Recently researchers from various cross-fields haveconcentrated in the field of DNA sequence analysis in order to extract the vast information contenthidden in it .Power Spectral Density of codingand non-coding regions of DNA sequences have been estimated by Parametric method and an attempt has been made to compare and differentiate coding regions from non-codingones. DNA (Deoxyribo Nucleic Acid) is a huge data base available to us in Public Domain havingheredity traits hidden in it . Genetic information is stored in the particular order of four kinds of nucleotide bases, Adenine (A), Thymine (T),Cytosine (C) and Guanine (G) which comprisesthe DNA molecule along with Sugar-Phosphate
backbone. There are two complementary DNAchains twisted around one another in a righthanded double helix structure. A straightenedform of a DNA helical structure Adenine (A) of one strand always pairs with Thymine (T) of opposite strand while Guanine (G)always pairs with Cytosine (C). DNA base sequence is always written from 5’ end to 3’ end which iscalled polarity of DNA chain.. Two types of nitrogen bases purine and pyrimidine are present in DNA.‘A’ and ‘T ’are purine bases whereas ‘C’ and ‘G’ are pyrimidine bases. The DNA strands are heldtogether mainly by

hydrogen bonds between bases. There are two hydrogen bonds between ‘A’ and ‘T’while three hydrogen bonds between ‘C’ and ‘G’. Hence ‘C’ and ‘G’ bonds are stronger than ‘A’ and‘T’ bondsThe DNA sequence can be divided into genes and inter-genic spaces. The genes can again besubdivided into exons (coding region) and introns (non-coding region). Even though all the cells in anorganism have identical genes only a selected subsets are activated in any family of cells. Exons of aDNA sequence are the most information bearing part because only the exons take part in proteincoding while the introns are spliced off during protein synthesis.Gene prediction refers to detecting locations of the protein coding regions of genes in a long DNAsequence.

There has been great deal of work done in applying Digital Signal Processing and Statisticsmethods to DNA .The authors have presented methods for identifying coding in recent past, some of which are mentioned here in a nutshell. It has beenestablished that base sequences in the exon regions of DNA molecules exhibit a period-3 property
because of the codon structure involved in the translation of nucleotide bases into amino acids.Investigation into long range correlation has also been the focus of attention for manyresearchers . A coding measure scheme employing electron-ion-interaction pseudo-potential(EIIP) was presented as a revision for binary indicator sequences.Implementation of digital filters toextract period-3 components and effectively eliminate back ground 1/f noise present in DNA sequencehas given good results .Positional Frequency Distribution of nucleotides has also giveninteresting results . In this article andnon-coding regions of DNA sequence based on graphical representation of PSD plots using low orderAuto Regressive Yule Walker Algorithm.
There is vast genomic data available in the NCBI Gene bank and DSP can be used as an effectivetool for analysis of this data. DSP technique is applicable only to numerical data but genomic dataconsists of four alphabets A,T,C and G. Hence a mapping technique is required to convert the 4-letteralphabet sequence into numerals before applying DSP techniques. Different researchers have adopteddifferent mapping methods for this purpose. Here the authors have attempted applying a new mappingrule based on weak-strong hydrogen bonding for digitization. As nucleotides ‘A’ and ‘T’ have twohydrogen bonds in their molecular structure they have been treated as weak bond and assigned integervalue ‘2’. Nucleotides ‘C’ and ‘G’ have three hydrogen bonds so they are treated as strong and have been assigned integer value ‘3’.
For example a DNA sequence of length N:
x[n] = [A T G C C T T A G G A T] (1)

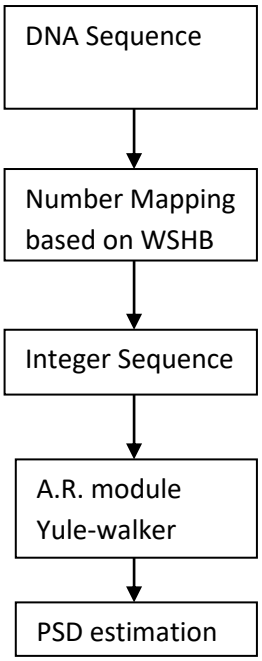
After mapping:
xsw[n] = [2 2 3 3 3 2 2 2 3 3 2 2] (2)
This method is employed to the data sequence for parametric analysis ofDNA sequences which models the data as output of a linearsystem driven by white noise and attempts to estimate parameters of this linear system. The mostfrequently used linear system model is the all pole model, a filter with all of its zeroes at the origin onthe z-plane. The output of such a filter for white noise input is an AR process, known as AR method ofspectral estimation. There are different types of AR methods such as Burg method, Covariance andModified Covariance method, Yule-Walker (auto-correlation) method etc. The advantage of Yule-Walker Autoregressive method is that it always produces a stable model. Parametric methods can yieldhigher resolution when the signal length is short .The output of such a system with white noise input referred to asAutoregressive (AR) process has been implemented here .

The pth order power spectrum of Auto Regressive process is given by:

Formula:
where b(0) and ap(k) are estimated from given data.

$$P_{AR}(e^{j\omega}) = |b(0)|^2 / |1+a_p(k)e^{-j\omega k}|^2$$

Flowchart



Block diagram realization of an AR model PSD estimation system.Here Yule-Walker Autoregressive method has been implemented efficiently for Parametric analysis of DNA sequence. AR models are popular because with them an accurate estimation of PSD canbe obtained by solving linear equations. Since in above equations |b(0)|2 is constant, the only valuethat is needed for calculating the shape of PSD are the coefficients ap(k). Though there are variousmethods to find these coefficients, the Yule-Walker (auto-correlation) method has been used here forits simplicity.It has been observed that Yule-Walker PSD spectrum plots are smooth, distinctand devoid of any spurious noise component. It is also evident from the plots that in case of Yule-Walker Autoregressive Power Spectrumestimation method the frequency resolution is independent of number of databases thatis applied to nucleotide databases. A graph of this power spectral density versus frequency is plotted for the genetic sequence in Homo sapiens(here for the defective genetic sequence chromosome number 11 and 16).This graph is decoded into human genetic sequence which determines whether the genetic structure of the amniotic tissue is healthy or not.

V. CONCLUSION

Before the birth of a child, testing of genetic sequences can be done.There is law regarding the below mentioned test that is genetic testing and screening which came into wide use with prenatal tests—amniocentesis.

A longstanding concern about genetic testing is that people at increased risk for a serious condition could face discrimination & sex determination of the embryo. This process of sex determination which has legal restrictions in certain countries prompted the passage of the Genetic Information Nondiscrimination Act in 2008.

This law is valid not only in United States ofAmerica but all throughout the world.

By analysing the situation the law should be modified & genetic screening via electronic machines& equipments should be allowed to eradicate genetic disorders for the benefit of the mankind .

As in many cases this could save a life if the foetus is diagnosed with a life threatening disease like Thalassemia.

ACKNOWLEDGMENT

I am grateful to my college faculties for encouraging and supporting me a lot for writing a paper on this topic.

REFERENCE

[1] Anastassiou D., “Frequency–domain analysis of biomolecular sequences”, Bioinformatics 16, 1073-1081.

[2] Anastassiou D., “DSP in genomics: Processing and frequency domain analysis of character strings,” IEEE,0-7803-7401-2001

[3] Ficket J.W. and Tung C.S., “Recognition of protein coding regions in DNA sequences”, Nucleic Acids Research,

[4]www.glown.com/..../Gamete%20 and %20 Embryo%20 Cryopreservation

[5]www.stjude.org/stjude/stjude/v/index.jsp?vgnextoid.
3)www.thalassemia.org/updates/BetaEnglish.pdf/v/index.jsp?vgnextoid.
[6] Li W. Kaneko & K, “Long range correlation and partial 1/f spectrum in a non- coding DNA sequence,” Europhys.
Lett., Vol.17, No.7, pp. 655-660, January1992

[7]Nair Achuthsankar. S. and Mahalaxmi T., “Are Categorical periodograms and Indicator sequences of genomes
spectrally equivalent?”

[8]Tuqan J. and Rushdi A., “A DSP based approach for finding the codon bias in DNA sequences”, IEEE journal on
signal processing, vol.2.No. 3, June, 2008.

https://en.wikipedia.org/wiki/Prenatal_diagnosis

[9]Chakraborty,NiranjanSpaniasA…LesmidisL.D&TsakalisK..”Autoregressive Modelling& Feature Analysis ofD.N.A sequences

[10] EurasipJournalon Applied Signal Processing 2004.I,13-28

[11] FickettJ.WandTung”Recognition of protein coding regions in D.N.A sequences.”Nucleic Acids Research.

[12] L.WKaneko&K.”Long range correleations in symbol sequences”.