DETECTING INSIDER AND MASQUERADE ATTACKS BY IDENTIFYING MALICIOUS USER BEHAVIOR AND EVALUATING TRUST IN CLOUD COMPUTING AND IOT DEVICES

A Dissertation
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Krishna Kanth Kambhampaty

In Partial Fulfillment of the Requirements
for the Degree of
DOCTOR OF PHILOSOPHY

Major Department:
Computer Science

May 2019

Fargo, North Dakota

North Dakota State University
Graduate School

Title

DETECTING INSIDER AND MASQUERADE ATTACKS BY IDENTIFYING MALICIOUS USER BEHAVIOR AND EVALUATING TRUST IN CLOUD COMPUTING AND IOT DEVICES

By

Krishna Kanth Kambhampaty

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

DOCTOR OF PHILOSOPHY

SUPERVISORY COMMITTEE:

Dr. Kendall E. Nygard
Chair

Dr. Vasant Ubhaya

Dr. Jen Li

Dr. Limin Zhang

Approved:

June 28 2019 — Date

Dr. Kendall E. Nygard — Department Chair

ABSTRACT

There are a variety of communication mediums or devices for interaction. Users hop from one medium to another frequently. Though the increase in the number of devices brings convenience, it also raises security concerns. Provision of platform to users is as much important as its security.

In this dissertation we propose a security approach that captures user behavior for identifying malicious activities. System users exhibit certain behavioral patterns while utilizing the resources. User behaviors such as device location, accessing certain files in a server, using a designated or specific user account etc. If this behavior is captured and compared with normal users' behavior, anomalies can be detected.

In our model, we have identified malicious users and have assigned trust value to each user accessing the system. When a user accesses new files on the servers that have not been previously accessed, accessing multiple accounts from the same device etc., these users are considered suspicious. If this behavior continues, they are categorized as ingenuine. A trust value is assigned to users. This value determines the trustworthiness of a user. Genuine users get higher trust value and ingenuine users get a lower trust value. The range of trust value varies from zero to one, with one being the highest trustworthiness and zero being the lowest.

In our model, we have sixteen different features to track user behavior. These features evaluate users' activities. From the time users' log in to the system till they log out, users are monitored based on these sixteen features. These features determine whether the user is malicious. For instance, features such as accessing too many accounts, using proxy servers, too many incorrect logins attribute to suspicious activity. Higher the number of these features, more suspicious is the user. More such additional features contribute to lower trust value.

Identifying malicious users could prevent and/or mitigate the attacks. This will enable in taking timely action against these users from performing any unauthorized or illegal actions. This could prevent insider and masquerade attacks. This application could be utilized in mobile, cloud and pervasive computing platforms.

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my deepest gratitude and respect to my advisor, Dr. Kendall E. Nygard, Chair of Computer Science department, for his constant support and encouragement throughout the period of my graduate years. I cannot thank him enough for his insightful feedback and wholehearted support that have helped me greatly in completing this dissertation.

I would like to thank all my committee members: Dr. Vasant Ubhaya, Dr. Jen Li and Dr. Limin Zhang for their helpful comments, time and interest. I sincerely appreciate their understanding and support.

I would like to thank Mr. Curt Doetkott and Qian Wen for helping me with data generation. I'm very thankful for Curt's help in the eleventh hour.

My family has been my backbone in my journey. Their support and constant encouragement have been my motivation factor to stay on course. Their belief in me has been my good reason to continue my dissertation journey. I'm what I'm because of them. Being grateful is the least I can do.

Just as Socrates said, 'if you get a good life partner, you will be happy, else a philosopher'. I'm grateful to have a spouse who stood by me while pursuing this degree. There were many times when research took priority. Despite this, there was only encouragement and support all the time.

I like to take this opportunity to thank my grandparents, maternal uncle, aunt and Venkat. I also would like to thank my friends Ragesh Bondada, Angshu Kar and Minhaz Chowdhury for their encouragement. Finally, I would be remiss if I did not mention that I'm thankful to all my friends, who have been constant source of support to me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

1. INTRODUCTION

The last decade has witnessed a phenomenal growth in the area of the wireless industry. This surge is both in the number of users as well as the service providers. It is anticipated that the mobile traffic will grow leaps and bounds by the year 2020. The expected networks connections will be around fifty billion [1] [2] [3]. This is nearly ten thousand times the current mobile traffic. This is primarily due to the power of evolving mobile technologies with always available connectivity, fast and reliable connection while on the move. Mobile communications have revolutionized our way of communication [4]. The advancement from 1G to 4G technology has been relatively fast.

Wireless devices have become part of our lives. It is like a woven fabric in our daily lives. Pervasive computing and wireless networking have been in existence for over two decades. Pervasive computing is also known as ubiquitous computing. Although there is a subtle difference in pervasive and ubiquitous computing. Being present everywhere is ubiquitous while being part of everyday activities of our lives in pervasive computing. Pervasive computing became part of our daily activities which minimizes the usage of computers. These are connected to internet/network and are available all the time. Unlike desktop computers, ubiquitous devices are available in any location, format and any device. These devices can be computers, laptops, refrigerators, wearable devices like smart watches, mobile devices like cell phones etcetera. Thanks to the advanced technology in sensors, microprocessors, operating systems and middleware in ubiquitous computing.

Ubiquitous computing was invented by Mark Weiser at the Olivetti Research Laboratory in Cambridge, England [5]. An employee ID card with a chip on it was first created to track the location of an employee in the building.

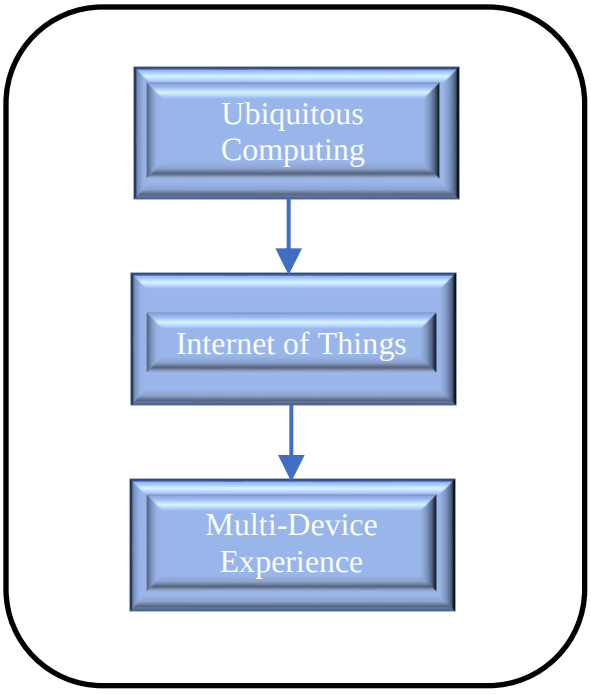

Figure 1. Pervasive Computing

Internet of Things (IoT) are a range of physical devices which work together to collect, connect and exchange user data. These devices surrounding us will be connected to a network. Sensor networks and radio frequency identification (RFID) play a major role in the IoT devices. IoT devices are smart devices that exhibit smart connectivity with network resources. The 4G LTE and strong Wi-Fi ranges helps the IoT devices to connect ubiquitously. Cloud computing can provide services to the IoT devices. Cloud computing can store the data received from IoT devices, do the necessary computations to analyze and interpret the data. This data can be translated to human understandable visualization. All of this happens in the background in an uninterrupted manner.

Consumers switch from one device to another and this happens seamlessly. Security has become a major area of research. Our research focuses on the security aspect of the IoT and Cloud Computing environment.

1.1. Significance of the Research

The dynamic user population in pervasive computing system suggests that the infrastructure used by users has to friendly to current and new users [6]. Inside threats [7] [8] originating from nodes that are compromised and malicious outsider attackers compromise the user identity by controlling the user's computing devices.

As the user population in cloud and pervasive computing is so dynamic and constantly changes, managing user profiles in cloud and pervasive computing environment has become difficult. As the pervasive computing environment is supposedly user-friendly, it also attracts malicious users who pose as new users to the system.

Genuine user identity is a necessary step towards building trust. Failure to correctly identify users can lead to drastic results of trust management. Traditional intrusion detection systems can detect attacks after the incident happened. However, prior to the attack, users might have exhibited certain behavioral patterns which could indicate the malicious behavior.

It is also well understood that any device that connects to internet faces the challenges of security and privacy. To realize the full potential of these devices, need to efficiently connect to secure networks. These are not impermeable to security vulnerabilities.

The first step in security is usually authentication. Authentication mechanism alone is not enough, as a security mechanism as identities can still be stolen. Other security mechanisms like introducing biometrics to authenticate might also not entirely suffice. In addition, not all devices can have this feature. Thus, trust management is necessary for utilizing the resources securely.

Computational trust [9] was introduced to use in addition to the traditional security mechanism. It depends on the trust of provider on requester. Many applications were built utilizing this trust mechanism. However, this kind of trust might not applicable in mobile and pervasive

computing as some of these devices do not depend on the service provide for sensing and transmitting information.

Pervasive computing systems are the next generation of computing systems. Pervasive computing environment facilitates users to hop from one device to another seamlessly. The dynamic nature of pervasive computing raises issues of vulnerabilities and security concerns. It is much more challenging than the traditional communication mediums like internet. One of the challenges faced by pervasive computing environment is the ever-changing and dynamic population, insider threat and masquerade attacks.

Hence a powerful user behavior management system needs to be incorporated in the security mechanisms. User behavior analytics as defined by Gartner is a, "cybersecurity process about detection of insider threats, targeted attacks and financial fraud". User Behavior Analytics (UBA) when incorporated along with traditional security mechanism can aid in elevating potential threats. UBA monitors users' behavior patterns while using the system to analyze and detect any anomalies in behavior. User behaviors such as types of applications launched, accounts logged into, types of services utilized etc.

In this dissertation, we are focused on designing and evaluating a user behavior trust management system. When abnormal behavior detected, rapid actions can be taken in order to stop malicious users from attacking the system.

We propose a solution that utilizes user behavior operations as the key to detect insider attacks as well as masquerade attacks. The goal of our model is user tracking to detect any anomalies. The monitoring of users starts right from the login until users' log out of the system. We have incorporated sixteen different user behaviors. Eight of them from user login behavior features and eight from operations behavior for our model. These features are used to analyze user

behaviors and detect anomalies. Once anomalies are detected, the trust value of users is decreased. After this, we classify users into six categories of trustworthiness. With the user classification, service providers can take necessary measures to mitigate the effect. This model can be tailored according to the requirements of the service providers.

1.2. Organization of the Dissertation

Chapter 2 presents the security concerns currently faced in cloud and IoT environments. It describes the significance of user trust in the security mechanisms. Also talks about related research work.

Chapter 3 describes in detail the problem statement and the importance of this research.

Chapter 4 presents the proposed research solution. Presents the detailed model view along with its design and development.

Chapter 5 presents the results obtained from our model.

Chapter 6 Finally, this chapter presents the Conclusion.

2. RESEARCH WORK

2.1. Mobile Networks

In lieu of the immense growth of mobile network communications and as a result its traffic, there is a lot of stress in the current and future mobile networks in an unprecedented way. The limited frequency spectrum poses a challenge due to increased mobile traffic as well as stringent requirements for 4G and 5G. There is also the problem of managing mobile traffic in heavy traffic loads (or hot cells), peak hours, network congestion due to gatherings, a sudden spike in large volume of calls in a particular location etcetera resulting in dropped calls, and possible system failure.

One of the low-cost alternatives is *load balancing*. Load balancing plays an effective role. It transfers the load off of the overloaded or congested network to the neighbors with free resources in order to enhance the network performance and overall user satisfaction. Lately, it has attracted quite a bit of attention due to its promising solution [10] for higher resource utilization, enhanced system performance and lower operational and maintenance cost. Figure 2 is a representation of load balancing which utilizes the resources according to the traffic in the network.

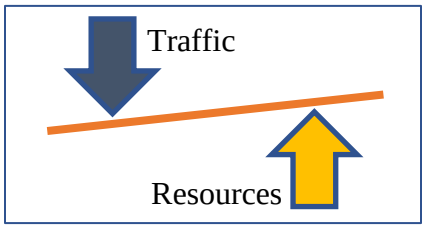

Figure 2. Load Balancing

Camellia Askarian et all propose an adaptive handover time scheme [11]. BS1 receives signal strength which is below specific threshold, this implies handover is necessary. When BS1 clarifies that BS2 will become hotspot due increased call load. BS1 initiates HOT-SPOT alarm to BS2. Any calls to BS1 and BS2 after this point are temporarily postponed. BS2 will be the hotspot

and execute all potential handovers earlier than what has been scheduled through conventional hand over mechanism.

2.2. User Trust in Cloud and Pervasive Computing

Cloud computing architecture provides various resources without the need to acquire them. Services such as storage, servers, computations, operating system, applications etcetera are provided in cloud environment. It provides a low-cost environment with scalability to industries. It has revolutionized the IT industry.

The main concern of cloud computing environment providers and users is about the security and privacy. The impact of damage in the cloud environment is far worse than users utilizing internet as a medium of sharing resources. Trust is an important factor to be addressed in the cloud computing.

Trust is a central component in human interaction. In the area of computer science, the word ‘trust’ is a widely used term. The definition of trust differs among researchers and application areas. Trust is one of the most influencing factors when it comes to human interaction. Without trust, the survival is difficult for families, houses, politics, friendships, relationships, markets etc. Trust can be considered as social glue that enables us to interact with one another. Humans cope with uncertainty using the trust mechanism.

Human trust with machines and Artificial Intelligence might not be that much different from trusting humans. In the AI context, there might be some reasons as to why trust has become popular. Lives saved in the military as a result, of robots replacing humans in a highly risky situations or autonomous driving vehicles etc. Trust plays a vital role in the acceptance of technology and the products. It continues its impact on businesses and economic behavior such as digital assistants like Amazon Alexa, Google assistant etc.

Trust is defined as [12] ‘confidence in or reliance on some quality or attribute of a person or thing, or the truth in a statement’. Trust management was introduced by Blaze M. in 1996. This was first implemented in cloud computing environment as a way of solving security problems. The word ‘trust’ is a complex and an abstract word. It is hard to pin point what comprises of trust as it is a multi-dimensional and multi-faceted. Trust is often built upon past or historical experiences of an individual. Online trust has few characteristics [13] that needs to be taken into consideration.

1. Trustor and trustee: There are usually two parties involved where trustor is a consumer and a trustee are a service provider such as cloud service provider or cell phone service provider etc.
2. Vulnerability: This comes into play when the environment is uncertain and risky. In an online environment, there is so much unpredictability from the service provider and consumers face the consequences and risks.
3. Produced actions: Trust developed over a period leads to actions which are risky. Users exhibit these behaviors when shopping online or in store.
4. Subjective matter: Situational factors and individual differences are directly impacted.

Recent study has described trust as fundamental construct for understanding users perception of technology. Initial trust formation is vital in overcoming the perceptions of risk and uncertainty before the usage of a novel technology.

There are three levels of trust [14] Inductive trust, Social trust and Moral trust. Inductive trust is experience which is derived from a person’s past experiences. People trust something or someone as they have acted as expected. Inductive trust is the simplest of the trusts and is easy to formalize. The second kind of trust is Social trust. This trust is dependent on the encounter between machines and humans. Machines have their own of set of goals just as humans. To trust or not

depends on the reasoning of humans. The third kind of trust is Moral trust which is based on sense of what is morally right. This kind of trust is hard to interpret and is the least explored area with in AI.

2.3. Relationship between Trust and Technology

Trust factors can also be utilized to improve the cyber security. Both the public and private sectors fall prey of cyber-attacks. Cybercrimes have costed the world nearly $3 trillion in 2015. This figure is expected to raise to $6 trillion by 2021 [15]. The cyber-crimes range from damage to data, loss of productivity etc. [16].

Trust has been playing a vital role in the domain of technology [17]. The usage of technology is directly related to the aspects of trust. Once, people start trusting the new technology for instance, its usage and sales drastically improve. Higher the trust, higher the usage of technology.

Security measures to tighten the cyber incidents only overwhelm the users. One of the studies from [18], there is a resistance from users in changing passwords. A survey comprised of 571 respondents from various walks of in life in the campus. Some were undergraduate students, graduate students, staff, researchers, faculty and administrators at Virginia Tech. When the passwords were required to be changed, only a portion of the individuals changed the password. In addition, the resistance to change password, changed from ‘rather not resistant’ to ‘strongly resistant’. Researchers of the study found that the even when the passwords were changed, it was perceived as unnecessary interruptions and were intentionally delayed. Study also found that password breach can attribute to security risks, it did not affect their attitude.

Raking up cyber security measures requires actions at many levels, starting from the design of the technology till its implementation and maintenance. Behavioral science can address the

cyber security issues. Sasse et all suggested that [19] security systems need an understanding of behavioral science. This prevents users from being the weakest link. Shari et al have done a survey on how behavioral science can impact the cyber security [20]. They describe that incorporating behavioral science into cyber security can yield to effective security system. They have focused their survey on two aspects: cognitive load and bias. Their survey suggests that including human behavior can lead to potential improvements in the cyber security system.

2.4. Trust in the Information Technology

Trust is emerging as a central aspect in the acceptance of Information Technology. The importance of trust in IT is more important than ever. IT has become center of our lives. We rely so much on the IT. For instance, an online reservation system, communication, hospital management, online shopping etc. all rely on IT. The trust in IT is not very different with the trust in people [21]. Trusting in humans means is belief in a person's capability to fulfill a task or a responsibility. Trusting in Information Technology means, it is understood as the system is functionally capable to complete a task to be done. Differences come in the aspects of integrity, morality. These are harder to be described in IT. There are several advantages of trusting in IT. Trust in IT influences the adoption or change to new technology and secondly it may affect risks, beliefs and attitudes to using a technology. Trust in digital environment is called as 'e-trust'.

2.5. Trust in the Cloud Computing Environment

Cloud computing provisions shared computing resources. Resources like storage space, servers etc. This is widely used in several industries such as healthcare, banking and education [22]. However, general public still do not have complete confidence and trust in the cloud computing. This is a might not seem as much of a problem at a first glance. This problem can

manifest into multiple folds. Industries are gradually getting more reliant on the cloud services for cost effective measures.

Cloud computing provides different types of services, such as Infrastructure as a Service (IaaS), Software as a Service (SaaS), Platform as a Service (PaaS) etc. Cloud Service Providers (CSP) offer these resources to users. The CSP's infrastructure is more powerful than the personal computing platforms. Though is true, cloud computing environment constantly faces security challenges. Some of them are Amazon S3 unavailability [23], iPad breach of personal data [24], President Obama's Twitter hack etc.

With the users not trusting the cloud, might affect the industries. This could also affect the cloud service provider industry. This user concern might have emerged from either misconceptions or lack of understanding of the technology. This misunderstanding on the cloud could have negative effects on the trusting the cloud. It might also be due to the security concerns of the data. If the understanding is that data is stored on a third party's hard disk, users would be worried about the encryption of the data or that someone can access the personal or confidential files [25]. In addition, the thought of relying on the cloud for data retrieval might seem like a problem to many. This could arise due the unreliable internet connection one might face. When users see the value for their money, the perception towards cloud computing might drastically improve. In addition, users have serious concerns over security and privacy of cloud computing.

2.6. Trust in Artificial Intelligence (AI) and Machine Learning

Artificial Intelligence has become ubiquitous in our lives. It is a common site among online shoppers to notice 'recommendations' from the website. Some of the online shopping sites such as Amazon, Google, Walmart have employed techniques like AI and Machine Learning to better understand the user requirements and entice them with the available products. Airline ticketing

sites, ride sharing services like Uber, personal digital assistants like Siri, Alexa, Google have relied on AI and Machine Learning technology to improve on the quality of service.

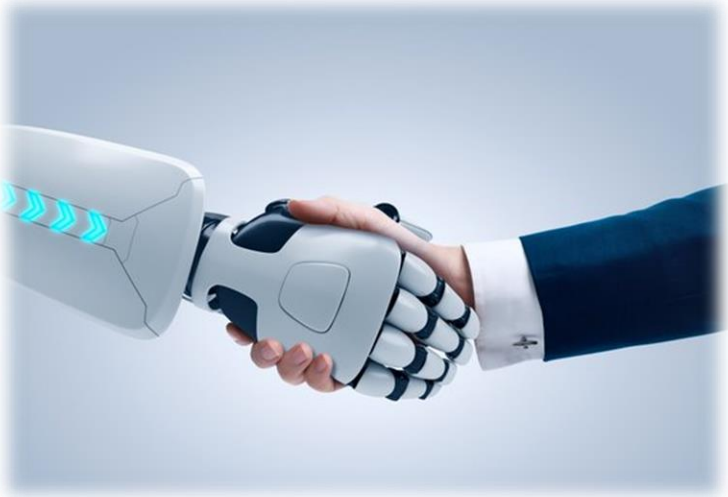

Figure 3. Trust and Technology go Hand in Hand
(Source: Getty Images)

AI is a software dominant technology and hence is prone to vulnerabilities. With this fact in hand, how much should the users be trusting on the outcomes generated by AI [26]. Autonomous cars are equipped with auto-pilot capabilities. This system can adhere to the rules of the road. One of the Tesla car has reportedly saved the driver during his regular commute. The driver upon realizing he had myocardial infarction, changed the destination to the nearest hospital. The autonomous car has efficiently maneuvered in the traffic and has safely brought the driver to the hospital. Upon arrival, driver was rescued by the hospital personnel. This scenario builds trust in autonomous vehicles and more specially in AI. However, in another scenario, a driver was killed in an autonomous car. Evidence suggests that the visual and radar systems have encountered a glitch, as a result, causing the death of the driver. In this scenario AI based system turned deadly.

John Launchbury, director of DARPA's Information Innovation Office, credits statistical learning as a second wave of Artificial Intelligence. This kind of learning has strong suite of learning but lacks in the ability to reason. The outcomes are sometimes skewed. This stresses on

the testing of machine learning to ensure the code running behind is flawless. John Launchbury suggests that the AI still needs to be perfected [26].

2.7. Build and Improve Trust

The question that remains is, how do we replicate the human to human trust to human to machines. Peter et al suggests that explainable AI can play a key role in establishing an initial trust on machines. This could also repair the trust relationships. When a machine such as an autonomous vehicle is providing explanation of its actions, humans can slowly gain trust as it reduces the perception of risk. This is key especially in the early stages of building trust. An explainable trust can also repair the broken trust. If an explanation is provided as to what caused the system to fail the expectations of its user, can also repair the trust.

The relation between explanation and trust [27] has been analyzed by Wolter Pieters. This analysis was accomplished using system theory and actor-network theory. Pieters made a clear distinction between trust and confidence. Confidence can be high, but the trust could be low. For instance, when the government announces that voting machines are secure, it builds confidence but might gain trust from people. There is also a distinction between explanation in Information Security versus explanation in AI. Pieters findings confirm that the explanation and trust especially e-trust are critical in the digital environment.

The next question that arises is how this explanation needs to be given. An explanation that could be understood by humans is required. For instance, people request for an explanation of decisions made by others. A series of verbal or pictorial explanation can easily make one understand. Machines need to closely follow this style of explanation.

Deception is detected with the identification of anomalies in environment [28]. The concept of trust has caught the eye of researchers in the past two decades. There has been a lot of research

conducted in this area in different disciplines from economics to medicine to information technology to data science. Trust has been part of various disciplines. It has been used in the electronic commerce, cloud computing, internet of things, ad hoc networks etc. The concept of trust lately has gained momentum. It comes into play when the traditional network security mechanisms fail such as access control, firewall, anti-virus etc. Conventional intrusion detection can only detect when the attack is happening but cannot prevent one. With user trust value already in the system, some of the attacks can be stopped. With malicious user detection, trust amongst genuine users is enhanced and overall system security is improved [29].

Li Wen et all proposed novel trust model [30]. In this model direct and indirect trust were proposed. First hand experiences are counted towards direct trust. This model combines both direct and indirect trust for user trust evaluation. User trust is vital as an inside attacker can compromise the system security. New users need to be assigned basic trust value. This way, trust is slowly built over time. This however has an obvious downside. Malicious attackers with bad reputation might re-enter the system as new users, thus gaining neutral or new reputation.

There has been a lot research done with respect to user identifications, by introducing either a secret key or a secret answer to prove individual's identity. There is still a flaw in this security mechanism, as identities can be stolen. Biometrics techniques can be applied for user identification. This is good mechanism to identify the user. However, all the systems where user has activity needs to have this feature. In addition, this would still not be able to prevent insider attackers.

Zhaoyu and Dichao [31] introduce a comprehensive system for detecting user behaviors in networks. In this system, malicious user behavior is stored in the database. They also propose a solution to identify suspicious users from re-attacking the system.

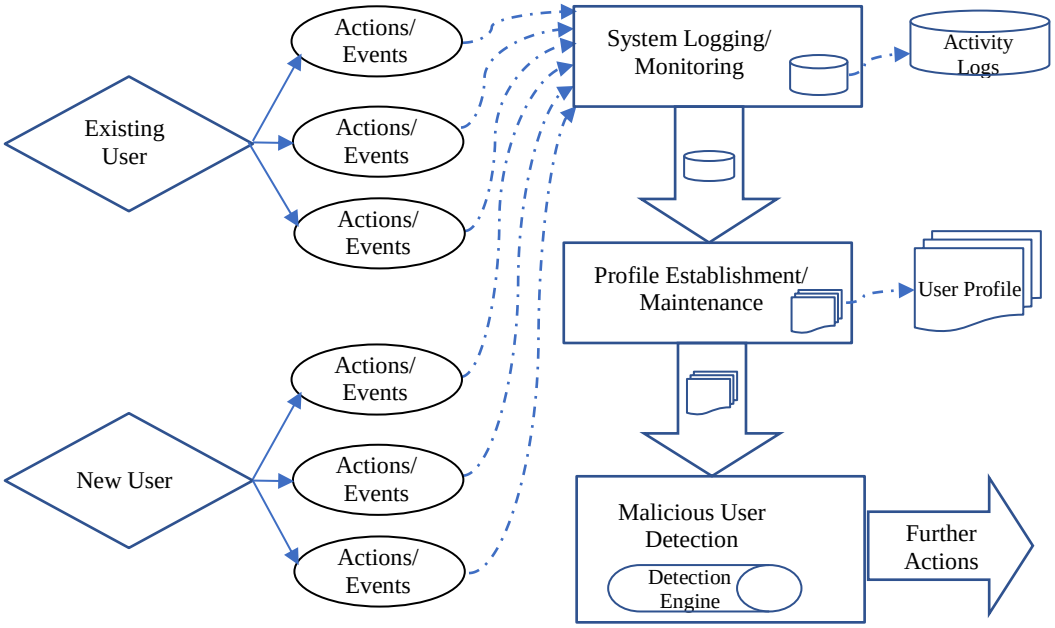

Figure 4. Overview of the Zhaoyu Model

2.8. User Behavior Trust Evaluation Models

Several user trust evaluation models have been proposed [32]. Some of them are:

1. Analytic Hierarchy Process (AHP) model
2. Fuzzy Mathematics based evaluation model
3. Role-Based Access Control model and other evaluation strategies

In the AHP model, trust evaluation is divided into three levels, namely Target layer, Property layer and the Evidence layer.

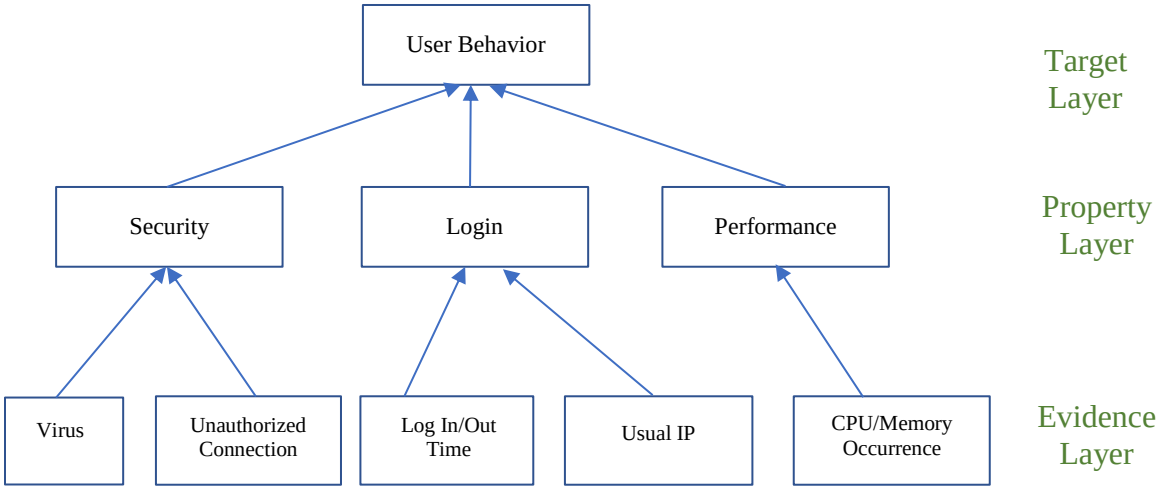

Figure 5. AHP Method

A dynamic trust evaluation model which combines Entropy and AHP were proposed by Jun et all [33] . The advantage of this model is the ability to balance objective and subjective weights to calculate the trust value. It can also calculate which user has consumed the highest amount of resources amongst other users in the cloud. However, this model suffers from obtaining user trust value for a short time period. It can only calculate for longer time periods.

Multi-level fuzzy comprehensive evaluation model [34] is a combination of qualitative and quantitative evaluation model. This model combined AHP with Fuzzy Comprehensive Evaluation (FCE). It can evaluate time impact principle. This principle calculates the number of times user has accessed the cloud. This model however, suffers from fraud risk problem.

Mohsenzadeh et all proposed a model based on Fuzzy Mathematics theory in Cloud Computing [35]. Trust evaluation is subjective due to several factors. This problem is eliminated using fuzzy mathematics. This model proposed two types of trust- direct and indirect trust. In direct trust, value is obtained from the local service provider whilst in indirect trust, value is obtained from other cloud service providers. This eliminates the problem of malicious users from hopping from one provider to another. However, this model also suffers from fraud risk problem.

A role-based access control model combined with user behavior trust has been proposed by Yang et all [36]. They proposed a multiple context to evaluate the user behavior. This model supports scalability and flexibility authorization strategy. This model just like other models described above suffers from fraud risk problem. It does not consider any of the evaluation principles. Also, this model is quite complex to be implemented in the cloud computing environment. It also lacks from measurement of trust.

Nathezhtha et all proposed an Improvised Long Short-Term Memory (ILTSM) model [37] for detecting masquerade attacks in the cloud. This model is an improved version of Long Short-Term Memory model. In this model, authors have used unsupervised machine learning algorithm to train and test data. The ILTSM model learns the behavior of a user and automatically trains itself and stores the behavioral data. This model can classify the user behavior as normal or abnormal. In addition, it can detect whether the anomaly node is a broken node or a new user node or a compromised node. Simulation is performed using Virtual machines and Cooja simulator. The ILTSM model was able to accurately predict masquerade attacks with 90% accuracy. However, this model uses unsupervised learning model and prior training and testing has been performed. Also, this model cannot detect Insider (or Malicious Insider) attacks.

Neural Networks based approach to detect suspicious user behavior was proposed by Martin et al [38]. Supervised Neural Networks was used to analyze the users' behavior and identify suspicious behavior. The training and testing were performed on simulation system and generated data. The accuracy obtained from their model is 98%.

Changping Liu et al [39] proposed an unsupervised method to evaluate user's network behavior and trustworthiness grades in a local network. Data collection was done for different network behaviors. The sample data was tagged with different trustworthy grades. Lastly,

according to the graded sample, support vector machine was constructed to evaluate user's network behavior.

Detecting a masquerader system in cloud computing environment is a challenge. Alguliev et all proposed this. This model is a two-step process. During the initial step, user profile is created and in the second detection is performed. User profile creation phase is further classified into two. User events log is recorded during the first phase. In the second phase, feature extraction is performed. After the user profile is created, the second step is the detection phase. In this step, cosine similarity method has been used. This is used to compare any abnormalities in the behavior with the stored normal user behavior. Any deviation is a good indication of the malicious intent. This is an effective model that can be easily implemented in the cloud environment. It can prevent masquerader attack. This model addresses the fraud-risk behavior. Few drawbacks are this model fails to identify behavior pattern and trust.

2.9. Security Threats

The growth of mobile networks, however has introduced a challenge, security. The security damages can range from vandalism, Denial of Service (DoS), identity theft, network spoofing, Man in the Middle, eavesdropping, malicious insider, traffic modification attacks, malware etcetera. These damages not only cause inconvenience to the users, but also loss of revenue for the service providers.

There are several kinds of attacks both on mobile devices as well as on the network [40].

1. Compromised Mobile Devices: When a mobile device falls prey to attacks, it becomes a compromised device. This device can be used to launch attacks on mobile networks. Mobile devices become compromised when they downloading applications from not genuine sources. When a device becomes compromised, attackers can steal the identity

of the users, utilize mobile device resources for malware attacks such as mobile botnets. These devices thus become botnet client, responding to the remote commands from remote servers.

2. Network attack: A typical network attack constitutes of utilizing eNBs (evolved Network Base) to access and attack Mobility Management Entity (MME) and finally taking down the entire network.
3. External Network: Mobile devices receive services from external or 3rd party providers. These can include roaming partner networks, internet browsing services etcetera. Users can be tracked, by pass authentication to gain free purchases etc.
4. Interface flooding: In this attack, interfaces in the mobile networks are attacked. These include radio and backhaul interface.
5. Eavesdropping: In eavesdropping, also known as sniffing or snooping attack, attacker(s) steal the information from mobile devices, network information. In this type of attack, unsecured network communications are the first to fall prey. Data packets being sent and received can be accessed. These attacks are hard to identify as no abnormality in the network communication is detected. Typically, eavesdropper(s) install a sniffer application either on the client or server to intercept the data. Public WiFi's for instance are very vulnerable as these are unsecured connections. As a result, sensitive and confidential information can be stolen during transmission. Intruder can also potentially manipulate the data and modify the network.
6. Unauthorized Data Access: In this attack, any data leakage can be accessed by intruder.
7. Data Modification: Data modification can occur when a flaw in the implementation is detected. This could also happen by exploiting network protocols.

8. Service theft: In this kind of attack, service is utilized without paying for it. Any authentication or authorization flaw could result in these types of attacks.
9. Malicious Insider: These are hard to identify as the intruder is from within the network. Users knowingly or unknowingly install or use malicious software in the computer or mobile devices.
10. Denial of Service (DoS): In this kind of attack, the target system is shut down making the host inaccessible to its intended users by cyber-attackers. This is accomplished by flooding the host system with traffic or sending information that leads to system crash. This makes the host system render useless to its legitimate users.
11. Distributed Denial of Service (DDoS): Distributed Denial of Service works similar to DoS with the only difference being, traffic flooding occurs not just from one source but from many different sources. This makes it hard to detect the origin and block the source of attack. This is often leveraged using botnet.

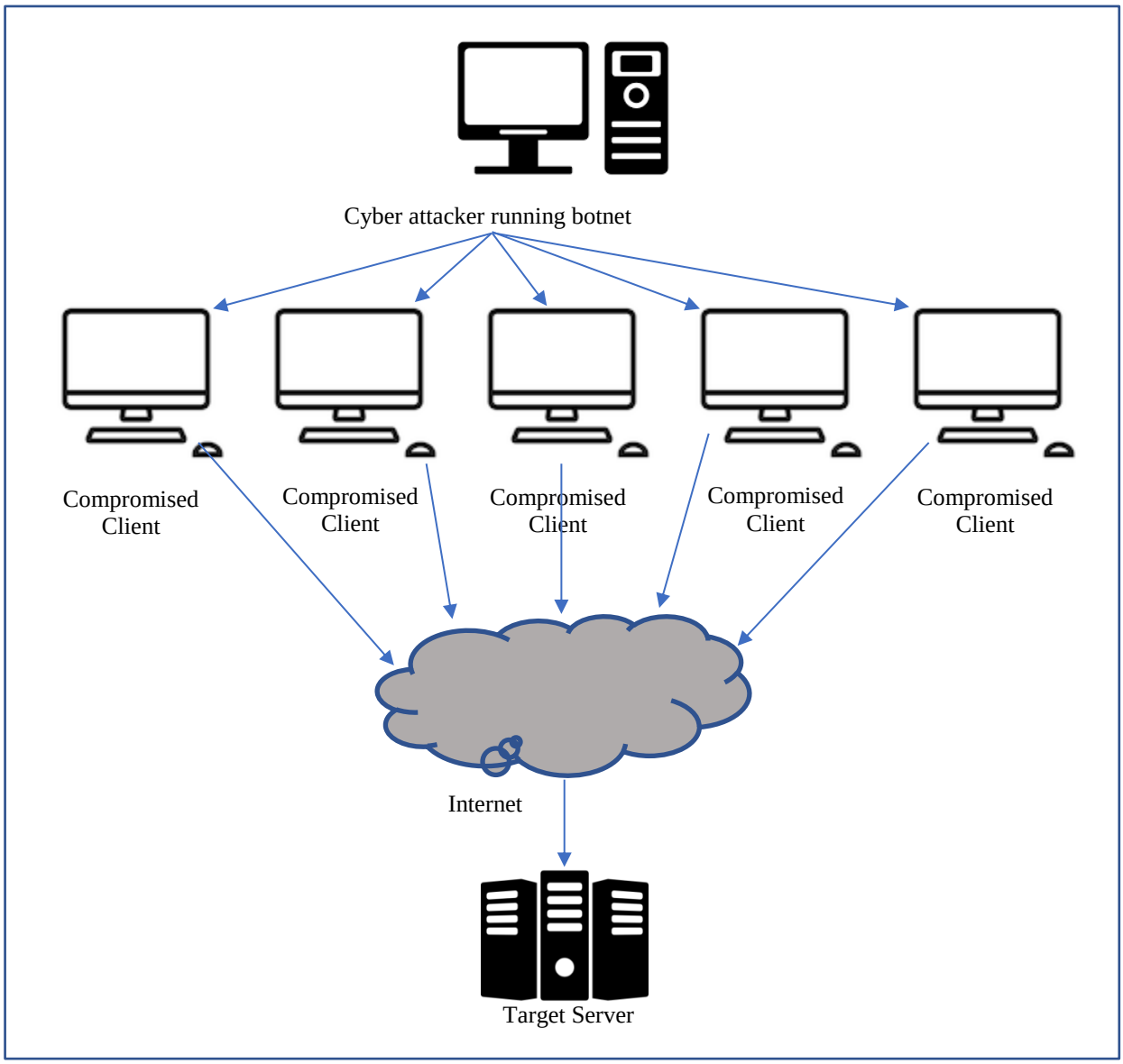

Figure 6. Distributed Denial of Service Overview

12. Malicious User: Apart from the above threats, there are several others. One of them is the malicious user attack. The damage that can occur because of malicious user is equally damaging if not more as compare to other type of attacks.

2.10. Principles for Evaluating User Trust

There are some principles [32] that need to be considered when evaluating user behavior. For our research we are considering the following principles.

1. Expired user behavior should not be considered while evaluating user behavior trust. These users should be treated as new users.
2. Abnormal or malicious user behavior plays a key role in trust evaluation.
3. When an abnormal behavior is detected, user is punished. Punishment is based on Rapid-Decline. The trust value of the user is declined quickly.
4. When users are found to repeat malicious behavior, their trust value is rapidly decreased.
5. Different types of evidences need to be considered before evaluating trust
    a. Security Evidence: Users present characteristics while using the resources. These are logged into each individual user's log file. Some of them are
        i. Traces of viruses
        ii. Scanning important ports
        iii. Any unauthorized connections
    b. Operation: Operating on the resources exhibits users characteristics. These are logged into user's log files. We consider the following
        i. Time spent by the user on the cloud
        ii. Functionalities usually used
        iii. Usually operation time on the cloud
        iv. Frequency of usage
        v. Data usage
        vi. Usage of another user's account
        vii. Any data definition or manipulation performed under different user's account

2.11. User Behavior Analytics

User behavior analytics as defined by Gartner is a, “cybersecurity process about detection of insider threats, targeted attacks and financial fraud” [41]. User Behavior Analytics (UBA) can help in solving potential threats. UBA keeps tracks of system user’s behavior patterns to analyze and detect any anomalies. UBA tracks user behavior such as applications launched and being currently used, accounts being accessed, services being utilized, network activity and most importantly files being accessed. File access such as any DDL (Data Definition Language) and DML (Data Manipulation Language) commands.

The goal of UBA is user tracking whether inside user or an outside user and detects any anomalies. Once these anomalies are detected, service providers can necessary measures to mitigate the effect. UBA is a close cousin of SIEM [42] where SIEM stands for Security and Information Event Management. SIEM finds the correlation among events that occurred in OS, networks, firewalls and other system logs using pre-defined rules. Though SIEM essentially focuses on logs, it is easy to miss the actual data itself. This is where UBA comes into play. According to Gartner, “User Behavior Analytics (UBA) is where the sources are variable, but the analysis is focused on users, user accounts, user identities and not on say, IP addresses or hosts. Some form of SIEM and DLP post-processing where the primary source data is SIEM and/or DLP outputs and enhanced user identity data as well as algorithms characterize these tools. So, these tools may collect logs and context data themselves or from a SIEM and utilize various analytic algorithms to create new insight from that data”.

2.12. Importance of UBA

1. Files or data are an important source for any industry. Leakage of sensitive data could be catastrophic. UBA models can alert any deviation in user behavior using the metadata and activity of its users in the system.
2. Service providers like cloud service providers, email service providers, storage service providers need to have a historical data of current users. Historical data such as user permissions, user accounts being accessed, access times etc. Machine algorithms once trained with such granular data, can differentiate between normal and abnormal behavior of users. Once an anomaly is detected, service providers can take necessary action to mitigate the effect.
3. Phishing attacks have become very common. When the users are not technical savvy, the passwords created are generally a form of their names with a sequence of numbers. Hackers exploit these kinds of users through phishing attacks and gain access to the system pretending as legitimate users. These masqueraders look like a legitime user to administrators monitoring the system and network. These hackers gain good control of the system by installing malware in the system which goes unspotted to anti-virus [42]. UBA can come to the rescue by identifying new behavioral patterns exhibited by the attackers pretending to be genuine users.
4. Perimeter-oriented security technology cannot prevent masquerade attackers or inside attackers. This kind of security mechanisms monitor network traffic and system viruses but completed miss masquerade attacks. With UBA model incorporated into the security mechanism get an edge on the attackers and prevent damage to the system, thus by saving time and money to the organizations.

2.13. Machine Learning Algorithms

Machine Learning is a scientific study of statistical and algorithms models [43]. Computers use machine learning to independently and effectively perform a task independently without the explicit instructions' requirement. Machine learning relies heavily on the input data. It learns patterns from the provided input data and infers the outcome. Learning experience among humans and animals is an innate quality. Learning from past experiences, pervious information and in learning in stages or phases is a process that comes naturally. This method of learning is imitated by machine learning algorithms.

Machine learning (ML) algorithms take sample data as input. This sample data is called as 'training data'. Using this training data, mathematical model is built. This training data trains the machine learning model in finding the patterns, relations, co-relations etc. in the data. Once this information is obtained, a portion of sample data, known as 'test data' is used to test the model. This test data helps us in obtaining the accuracy of the machine learning model. This accuracy determines how well the model has been trained. Higher the accuracy, better is the performance of the model. Another factor which contributes to the performance of the model is the sample data. The chosen model is only as good as the sample or training data. Training data needs to be closer to the real-world data.

The area of machine learning and statistics are closely related. Michael I Jordan, a scientist and a professor at the University of California, Berkeley, said that the mechanism of machine learning right from the methodological principles to theoretical tools, have had a long pre-history in statistics. Applications of machine learning are numerous [44].

1. Virtual Personal Assistants: There have been few virtual personal assistants available in the market. Siri, Alexa, Google Now are some of them. These devices are voice

activated and provide daily schedule information, weather forecast etc. Personal assistants collect information and refine this information based on the past interactions. This tailored information is later utilized to provide results to users.

2. Traffic Predictions: When vehicles use GPS navigations services, information such as current location and speed is saved in the central server. This information is used to manage traffic along with preventing traffic congestion. Ride sharing apps also use this information to estimate price of the ride and also assist in detours.
3. Security Video Surveillance: A single security person monitoring multiple video cameras is both a tedious job and at times boring. In addition, probability of missing an event or incident are high. These problems can be avoided, when a surveillance system is trained using machine learning.
4. Spam and Malware Filtering: This is one of the most common examples of machine learning. Some of the spam filters apply rule-based technique. In order to make sure that the new spam emails are filtered out, machine learning approach is used. Spam filtering is powered by machine learning algorithms.
5. Customer Support: In the past customer support offered by service-oriented companies had often had a live agent to cater to the needs of a client. However, this trend is slowing changing. There are chatbots on the other side of the line. These bots extract information from the website and present it before customers. Current chatbots also understand the basic user queries and provide answers.
6. E-Commerce Product Recommendations: Depending on a purchase, e-commerce sites also offer relevant product recommendations. This is tailored to each individual

purchaser. One of the famous examples is, an online retail giant figured out a customer was pregnant based on the customer's shopping patterns.

7. Fraud Detection: Machine learning algorithms are being used by banks and credit card companies to detect online fraud.
8. Credit Decision Making: Earlier loan companies used questionnaires to gather information about the financial background of the people who have applied for a loan or credit card. This questionnaire later was used in making decisions. This process was inefficient as 10%-15% of applicants fell in the borderline region which were then referred to loan officers. This process was rather inefficient. Machine learning was introduced back in late 80's. This improved the borderline applicants by 70%.

Machine Learning models have one common goal. It is, improving the performance of a task by exploiting patterns in the training data. Despite the similarities amongst machine learning models, Schlimer and Langley [45] have identified five main paradigms, each of the machine learning models fall into.

One of the major paradigms is associated with the field of neural networks. Knowledge is a multilayer network, which spreads from input to output nodes. Weights regulate the activation rate. A second paradigm is case-based learning or instance-based learning. Knowledge is stored in specific cases or experiences. These cases are cases are matched to new situations. The case matching relies on flexible matching algorithms.

A third framework is genetic algorithms. The representation of knowledge is in Boolean or binary format. It is an all or none matching process. A fourth paradigm is rule induction. Rule induction employs condition-action rules. One of the examples for this type is decision trees. Last paradigm is analytic learning. Knowledge is represented as rules in logical format. Learning

methodology utilizes background knowledge to construct explanations to solve similar problems with less amount of search.

The tasks of machine learning are classified into the two types. Supervised and Semi-Supervised learning: In supervised learning models, a mathematical model is built based on provided sets of data which consists of inputs and known outputs. The datasets contain training data with known samples of inputs and outputs. In the semi-supervised model, the training datasets might be missing some of the known outputs. Supervised learning algorithms, through the process of Iterative optimization of an objective function, model a function that predicts the output with the associated input. Classification and Regression are two models of supervised learning. In cases where the outputs are limited to a set of values, usually classification algorithms are used. Regression algorithms are used when the outputs have numerical values.

Unsupervised learning: In unsupervised learning models, the data sets do not contain known outputs to train the model. Algorithm finds similarities in the data and groups or clusters them. This implies that the algorithm learns directly from the test data as there is no training data available. The data in the test data is not labelled or classified.

Reinforcement learning: In reinforcement learning, the goal is to maximize the cumulative reward. Unlike supervised algorithms, where the datasets are labelled or classified, in the reinforcement learning model, the datasets need not be classified. It is employed by various software agents to find the best possible path to take in order to reach a specific situation.

In this methodology, we use three supervised learning models. They are

- Decision trees
- Logistic Regression
- Support Vector Machine (SVM)

- Random Forest

2.13.1. Decision Trees

Decision Trees: “A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility [46]. Decision tree algorithm only contains conditional control statements. These are trained on data for the purpose of classifying and regression problems. Decision trees are often fast and accurate algorithms and are commonly used. Decision trees have a flowchart like structure. It essentially has three main components

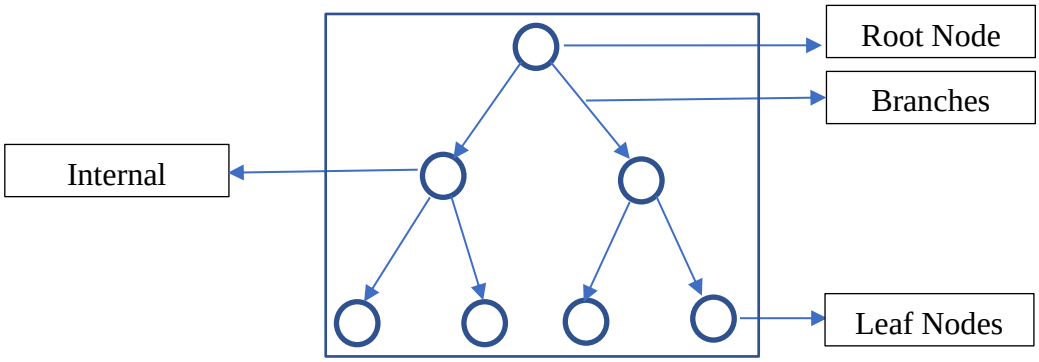

Figure 7. Decision Tree

- Internal nodes: These are the test values
- Branches: These represent the outcome of a test
- Leaf nodes: These are the labels, which is the outcome of a decision.

Decision trees learn from observations from the branches and resolution from the leaves. Models which can input discrete set of values are classification trees, whereas trees that can input continuous values are regression trees. In our model, the input set of values are discrete.

2.13.2. Support Vector Machine

Support Vector Machine (SVM): Support Vector Machine is a supervised learning model. It is one of the popular classification techniques. According to Hsu et al [47] SVM is considered

to be easier than Neural Networks in usage. Users not familiar with SVM, at first, do not often yield better results from this model.

The development of SVM algorithm has been accomplished in reverse order. SVM have evolved from sound theory while NNs were implemented using heuristic path. similar i to Neural Networks (NNs). SVM's have not been a popular tool for many in the beginning due to its strong theoretical background. They later became popular when the results from SVM are excellent in digit recognition, computer vision and text categorization [48] .

Computing the SVM classifier in the form:

$$\left[\frac{1}{n}\sum_{i=1}^{n}\max\left(0, 1 - y_i(w \cdot x_i - b)\right)\right] + \lambda\|w\|^2$$

2.13.3. Logistic Regression

This model is one of the popular statistical models [49], and uses logistic functions or logistic curve, to model a binary dependent variable. Logistic regression or logit regression estimates the parameters of a logistic model. A binary logistic model comprises of two variables, a pass or fail, head or tails, win or lose etc. These variables are represented numerically by '0' and '1'. Binary logistic regression has extensions. They are multinomial logistic regression and ordinal logistic regression. For our model, we limit our focus to logistic regression.

Applications of logistic regression model are in numerous fields. Areas like natural language processing, marketing, engineering, machine learning, medical fields etc.

General form of expression:

$$\ell = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

where x1 and x2 are two predictors,

l is log-odds,

and Bi is the parameters of the model

The corresponding odds are:

$$o = b^{\beta_0+\beta_1 x_1+\beta_2 x_2}$$

where b is the base of the logarithm and exponent.

The corresponding probability is:

$$p = \frac{b^{\beta_0+\beta_1 x_1+\beta_2 x_2}}{b^{\beta_0+\beta_1 x_1+\beta_2 x_2}+1} = \frac{1}{1+b^{-(\beta_0+\beta_1 x_1+\beta_2 x_2)}}$$

2.13.4. Random Forest Machine Learning Algorithm

The labelled data from both the tables (login and operations) are fed as input to the Random Forest machine learning algorithm.

Random Forest is an ensemble algorithm. It is a large collection of decision trees. Random forest is a best fit for classification of data, regression and other similar tasks that require multitude of decision trees. This algorithm is a multilayer decision tree. It corrects the problem of decision tree overfitting from training data and improves classification accuracy. Random forest was created by Tin Kam Ho [50]. The extension of this algorithm was developed by Leo Breiman, Adele Cutler which is currently owned by Mintab, Inc.

3. PROBLEM STATEMENT

The growth of mobile, cloud and pervasive computing has introduced security challenges. IoT and cloud environment are a classic example of distributed environment. The traditional centralized access control mechanism cannot satisfy security requirements required for these devices. User identity will also not meet security requirements. Only genuine users should be able to receive resources. Traditional intrusion detection system only detects after the attack has happened, but not before or during the attack.

Applying traditional access control methods in the cloud and pervasive computing alone cannot solve the uncertainty and vulnerability caused by the open conditions in these environments. Access control mechanism is one of the most important measures to ensure the security of Cloud and IoT environments. User authentication has been the front end of electronic communication for a very long time. It has been a binary event, whether the user gets access to an application or not. Hackers expose user credentials in dark web. Some new forms of authentication mechanisms such as biometric recognition, digital certificate, dynamic passwords, multi-factor authentication are relatively better than user credentials. But identity authentication cannot prevent malicious behavior of legitimate users or inside attackers and masquerade attacks. Therefore, evaluating trustworthiness of users is of great importance in security.

The research of trustworthiness measurement of user behavior is the hotspot in the domain of network security. Effective and accurate user behavior trust model can make the system respond well before the attack [51]. With the binding of security and reliability, data security can be effectively guaranteed. Therefore, building a trust parameter into security mechanisms can enhance the security. Users exhibit specific behavioral pattern which could imply an attack on the

system. If this behavior can be caught, users can be stopped from accessing the resources. Hence, this is an important research problem.

4. PROPOSED RESEARCH SOLUTION

4.1. Introduction

The exponentially fast growth of Internet has connected people across the world. Electronic commerce businesses flourished, information knowledge and sharing has become so much easier, endless supply of entertainment, internet of things, cloud computing and storage etc. One of the important factors for its rapid growth is due to its low cost of infrastructure. However, this rapid growth came with a price. Internet security has become a prime concern. The cost of cyber-crimes in 2015 was $3 trillion. This figure is expected to raise to $6 trillion by 2021 [16].

4.1.1. Internet of Things (IoT)

Internet of Things (IoT) is an extension of the Internet. In IoT, the connections extend to everyday used physical devices. These devices are embedded with system software and internet connectivity to communicate over the internet and can be remotely controlled. Smart phones, smart house, smart lighting, digital assistants, elder care applications etc. are just a few examples of IoT. However, just like Internet, IoT is no exception with respect to security. Security risks are higher as more and more people connect using IoT devices. One of the infamous examples is how a cyber-attack occurred on an IoT device. A fish tank in a home was connected to the internet. This connection happens to be unsecure connection. Attackers found this loop hole and were able to penetrate into the home network through this fish tank. An alert from Federal Bureau of Investigation (FBI), details that cyber attackers use IoT devices as proxies for anonymity and perform cyber-attacks [52].

The below Figure 8 gives an overview of Internet of Things. Some of these icons were used from www.flaticon.com

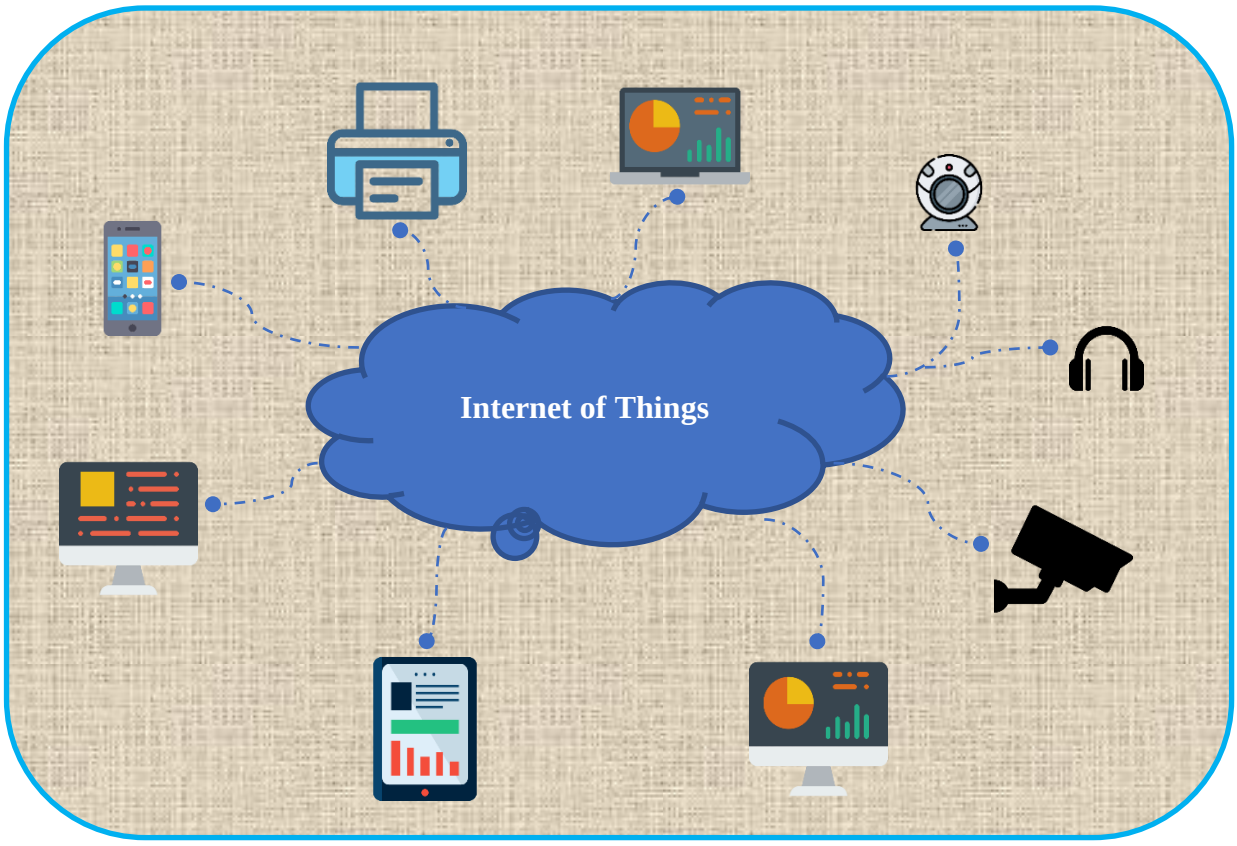

Figure 8. Internet of Things

4.1.2. Cloud Computing

Cloud computing architecture provides organizations with computing services such as storage, servers, services and applications without the need to physically acquire them [22]. Cloud computing provides three kinds of services.

1. Infrastructure as a Service (IaaS)
2. Platform as a Service (PaaS)
3. Software as a Service (SaaS)

In cloud environment, users have direct access to the resources provisioned by cloud service providers. Malicious users can damage the resources. Attacks such as masquerade or insider threats happen under the radar without the notice of anti-virus and/or system administrators. The below Figure 9 gives a bird's eye-view of cloud computing. This image is used from an online

source Wikimedia website. This is being used under the free to share license as stated on the website (https://commons.wikimedia.org/wiki/File:Cloud_computing.svg).

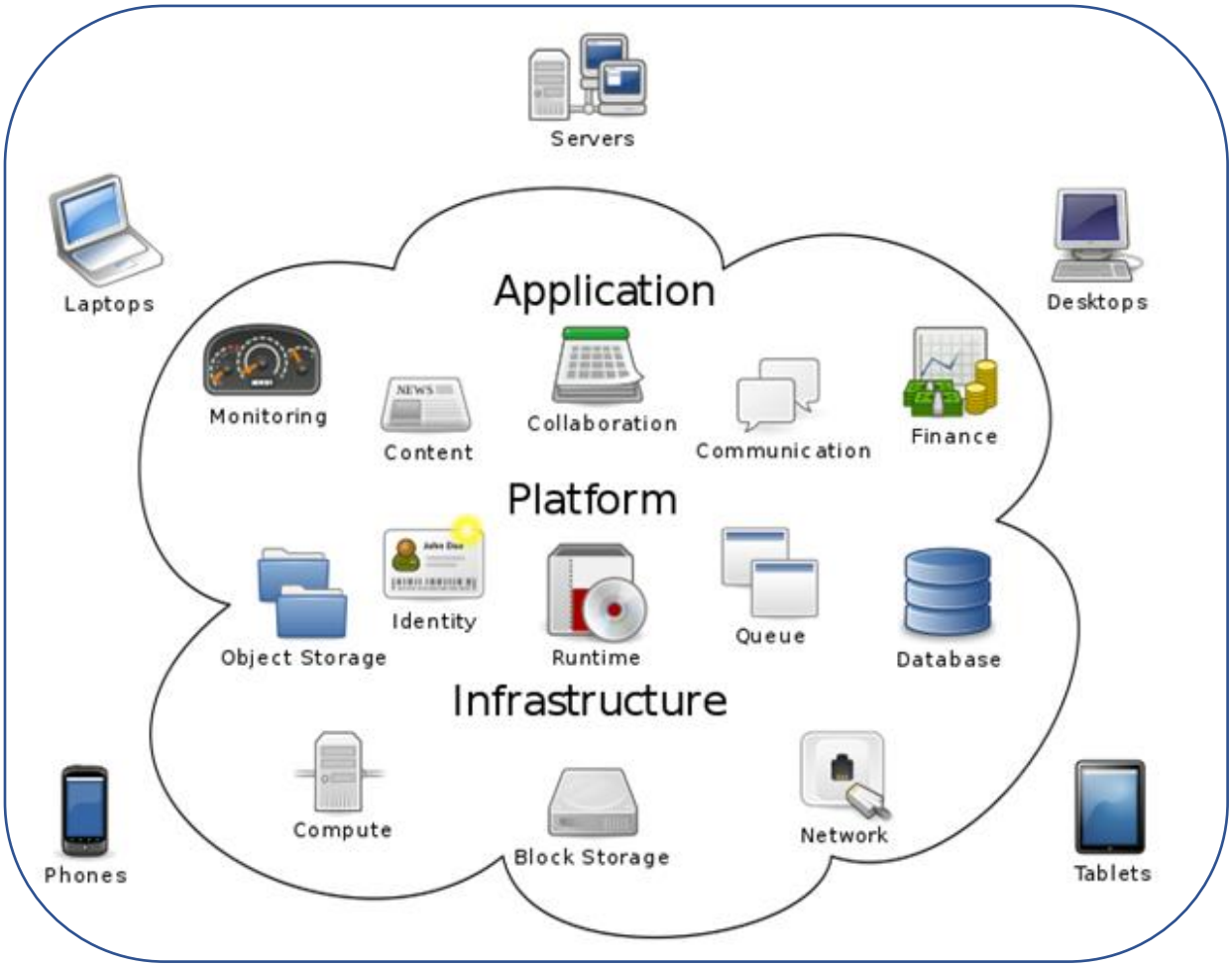

Figure 9. Cloud Computing Architecture
(Source: Wikimedia)

Security measures are in place to prevent cyber-attacks and damages. A lot of expensive software tools are used to detect and monitor the attacks. However, the attackers are increasingly using smarter ways to penetrate into the system and launch attacks. These attacks are coming from several different directions including inside and masquerade attacks. Security can be enhanced when user behavior and trust are incorporated into the existing software tools. Tracking and detecting malicious users can prevent attacks and mitigate the impact.

4.2. Proposed Model

Traditionally attackers focused on the corporate computers. When a server gets attacked, a host of computers in an organization are affected. The loss in time and money is huge. Organizations have spent over $86 million last year for security. In spite of this, two thirds of these organizations have still been breached. 80% of these breaches happen due to compromised credentials. Recent cyber-attacks have been geared towards individual accounts. This implies that when an individual account is hacked, attackers, effectively masquerade as that user and inherit that particular user's permissions. If this account happens to be an administrative account for instance, attackers have access to company's confidential files and data. Attackers could also meddle with the data or transfer the data to a different environment like a different location or different company etc. Malicious insider attacks have been reported to be nearly 25% of cyber-attacks.

To prevent and/or mitigate malicious insider, phishing and masquerade attacks, we proposed a solution. Plugging the user input data to our model, we will be able to classify users into two categories- genuine or ingenuine. In addition, each user is given trust value. This trust value represents how trustworthy is the user when using the system. This value ranges from 0 to 1. A zero value is the lowest trust and a value of one is the highest trustworthiness.

4.2.1. Goal of the Model

Given user behavior data from system logs, can we classify users into trustworthy or untrustworthy users? Our model offers a solution. From the system logs of user behavior data, our model can identify whether user is a trustworthy or not.

4.2.2. Birds Eye-View of the Model

In this dissertation we propose a novel approach which would not need any infrastructural or major software changes. Users exhibit behaviors when using the system. User behaviors like the number of attempts to login, logging into the computers at a certain date and time, from a certain place (IP address), accessing certain files etc. These user behaviors are the key to our model. The *login behaviors* and *operations behavior* are compared with normal users behavior. If an anomaly is detected, we have identified malicious users. The following Figure 10 provides the overview of our model.

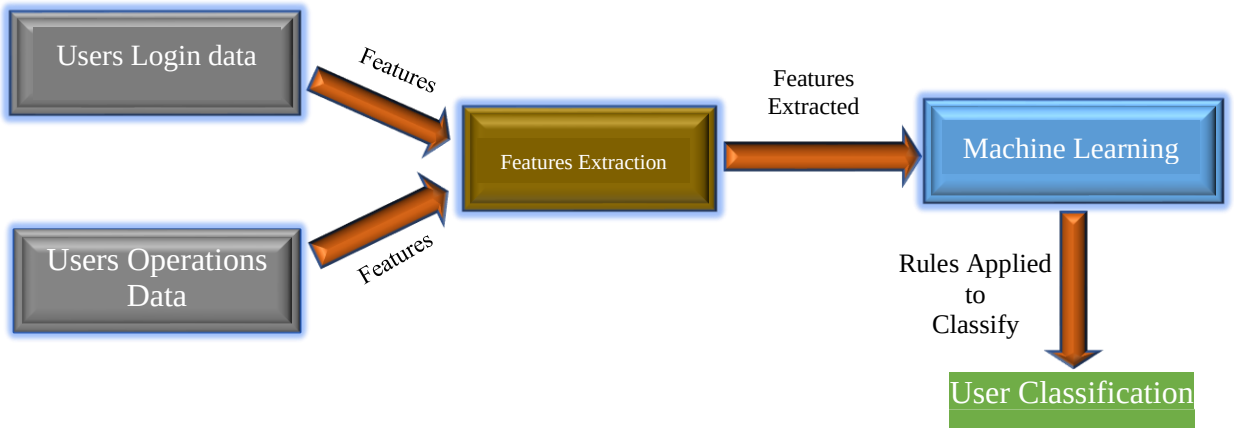

Figure 10. Bird's Eye-View of the Model

As we can see from the above figure, there are two sets of input data. One is Login data and the second is Operations data. Both of them are users' system logs right from the login until log out. Important features are extracted from the input data using Features Extraction process. After the extraction of important features, both the data sets are combined. The combined data set is used to train and test the model using machine learning approach. Fuzzy inspired rules are applied to classify users. The classified users were applied with trust value.

Login and operations data contain a total of six features. After doing feature extraction, the total number of features are twelve. These features data are used to train and test four machine

learning algorithms. Each machine algorithm's accuracy prediction is calculated. Best accuracy and the corresponding model are used to test the data. The accuracy from this model is taken for further classification. Inspired from fuzzy logic rules, we have written rules to classify each of the users into six different categories. These six categories form the output of the model. They are,

a) Very High Trust
b) High Trust
c) High Medium Trust
d) Medium Trust
e) Low Trust
f) Very Low Trust

With the usage, users' trust value is either increased or decreased. Trust value goes higher when the users do not exhibit any anomalies. Anomalies such as, deleting files without privileges, logging into the system from unknown IP address etc. When an anomaly is detected in the behavior, users trust value goes down. Lower the trust value, lower is the reputation of the user, leading to less system privilege or access or complete removal of user(s) from the system. The trust value varies from zero to one. Zero is the lower end of the spectrum and 1 is the higher end. User's trust value of one, is considered to be a trust worthy or genuine or legitimate user. If the users trust value is zero, that user is considered not trust worthy.

Identifying malicious users from their user behavior, triggers alarm to the service providers, who can take necessary action(s) to prevent further damage and alleviate the effect. This could prevent inside and masquerade attackers from damaging the system. This application could be utilized in mobile, cloud and pervasive computing platforms.

The following is the flowchart of the model.

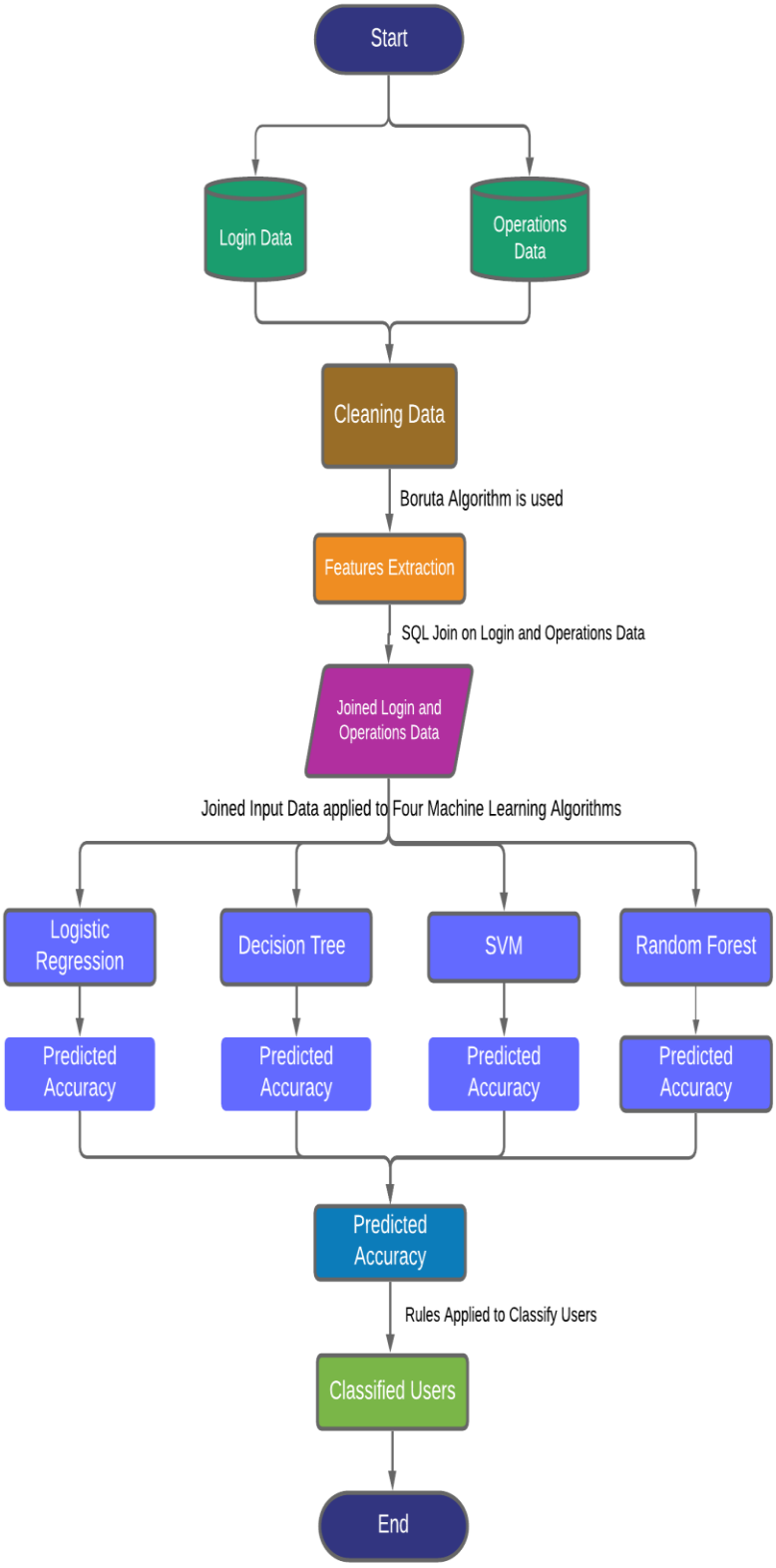

Figure 11. Flowchart of the Model

4.2.3. Detailed Functionality of the Model

The proposed model is divided into the following steps

1. Input Data
2. Login Data Table
3. Operations Data Table
4. Processing the Input Data
5. Features Extraction using Boruta algorithm
6. Combining the login and operations input data
7. Accuracy Predictions from four Machine Learning Algorithms
8. Choosing the Best Predicted Accuracy
9. Classifying Users into six categories

4.2.3.1. Input Data

Data required for the supervised learning models is very important, as this data is the key for training and testing the model. Input data plays a vital role in the accuracy of the model during initial training and testing phases. The importance of the data is that, it allows the model to accurately extract important features and find the relation(s) with the target feature(s). As our model classifies the users, the task at hand is to distinguish between genuine and ingenuine users. As we are using supervised model, features in the data need to be labelled so the model can learn the different features of the data. In addition, data used to train the model cannot be skewed. This implies that the number of genuine users and ingenuine users are equal in number in the training data. This equal ratio of genuine and ingenuine users helps the model to not lean towards one side or other.

Real-world data fetching was not possible for our methodology for several reasons. Firstly, we were not able to find data with many features. Some of the data found had fewer than 5 features. Our goal was to have at least 10-15 features in the data. Secondly, some of this data had way less suspicious users than the genuine users. This could be a problem, as the model might yield more false positives. This is will reduce the accuracy of the model during testing phases. Thirdly, getting data from multi-national companies was next possible due to the privacy and safety of its users. Due to these above conditions, we decide to simulate our data.

4.2.3.2. User Classification

The simulated data is divided into two types. Login Data and Operations Data. Each data has certain number of features required. Login data has 8 features plus a target column. Operations data has 8 features including a target column. The target column is a labelled column that is used to train and test our model. Target column is labelled as 'Group'. This column has two values. 'G' for genuine users and 'U' for ingenuine users.

Based on the column values, some users are categorized as genuine users and others are ingenuine users. There are few criteria's we used to distinguish between genuine and ingenuine users. Genuine users usually have a regular usage cycle. Majority of the times, they use the same computer, same location of login, similar user activities. Also, genuine users tend to log in to the system within few attempts if not the first time. We have taken these general behavior trends amongst users to classify them as genuine. We also considered about users travelling and/or use different devices (such as Windows or Mac). If the IP Address and Device are different from regular usage, these users might still be able to login with the few attempts. In addition, the Login Success is usually 'yes' (successful). Though the IP Address and Device values are different

factors being different from normal, other factors can aid in determining that the user is a genuine user.

The second value in the ‘group’ column is ingenuine users or ‘U’. Users that do not adhere to the principles of genuine users fall into the category of ingenuine users. Features such as User Proxy, High Accounts, User Login Attempts and IP Address are the triggering factors to categorize users as ingenuine. Of these features, High Accounts plays a critical role in deciding the classification of the users. When a user is trying to access high number of accounts, that particular user can be categorized as ingenuine.

4.2.3.3. Login Data Table

The login dataset has 20,001 records. Out of which, genuine and ingenuine users are divided equally in order to remove any bias during training phase. Genuine users are 10,001 while ingenuine users are 10,000 users.

Table 1. Login Data Features

ID	Login Date	IP	DEVICE	LOGIN SUCCESS	USER LOGIN ATTEMPTS	USER PROXY	HIGH ACCOUNTS

Login data has 8 features. Login data table has 20,002 records. Data comprises of 8 columns. The following Table 2 is a sample of the login data.

Table 2. Sample of Login Data

Group	ID	Login DT	IP Address	Device	Login Success	U Login Attempts	Use Proxy	High Accts
G	1	50.29494	1	4	0	1	0	0
G	1	57.59898	1	4	1	1	0	1
G	2	78.28549	1	2	1	1	0	0
G	3	85.84056	1	4	0	1	0	0
G	3	64.32878	0	4	1	1	0	1
G	4	70.88004	0	4	1	1	0	0
G	4	44.93313	1	4	1	1	1	0
G	5	38.79496	1	4	1	1	0	1
G	6	84.58626	1	4	1	1	0	0
G	6	71.14807	1	1	0	4	1	0
G	7	34.4708	1	4	0	3	0	0
G	8	86.48139	1	3	1	1	1	0
G	8	88.01191	1	2	0	2	0	0
G	8	71.01971	1	4	1	1	0	1
G	9	66.44137	1	1	1	1	0	0

a. User ID: This is the user ID column. Each user that is attempting to login to the system is assigned an ID number. The system can be a cloud environment or an IoT device. Repeat users will have the same ID as previously assigned. They will not be assigned new user ID. This is required to have a historical data of users. Data type is integer.

b. Login Date: The data has been generated for a span of 90 days. The numbers imply the day and time when the user has made an attempt to login. For instance, a date value of 50.5 implies that user has logged in on 50th day at noon.

Table 3. Login Date

Value	Date and Time
50.5	50^{th} Day at 12pm
1.5	2^{nd} Day at 12pm

c. IP Address: In this simulated data we have taken binomial values. For the sake of convenience, the IP values are taken as regions rather than the usual number format. Also, for our research, we have taken only two regions. Those values are USA and

abroad. USA implies that the user is from within the usual usage region. Abroad represents that the user is outside of usual usage region.

Table 4. IP Address

Value	Type
1	USA
0	Abroad

d. Device: The device column has four different values each representing a type of device. The column is a multinomial value and is of type integer.

Table 5. Device Column in Login Data Table

Value	Device Type
1	Windows Operating System
2	Macintosh Operating System
3	Linux Operating System
4	Mobile

e. Login Success: This column determines if the user attempt to login has succeeded or failed. Two binomial values represent the success or failure.

Table 6. Login Success or Failure

Value	Type
1	Success
0	Failure

f. User Login Attempts: This feature signifies the number of attempts users have made to log into the system. This is a multinomial value ranging from 1-4. When the users have made at attempt more than 4 times, those users will not be able log into the system after the fourth attempt. Those set of users are not considered further.

g. Usage of Proxy: This represents whether users have used any proxy servers to attempt to log into the system. This is a binomial value. Table 7 provides the input values required for proxy servers.

Table 7. User's usage of Proxy Severs

Value	Type
1	Proxy Usage
0	No Proxy Usage

h. Accessing High Number of Accounts: This is also a binary value, Yes or No. These values signify if a user has been trying to access more than one account at the same time which they have not done in the past.

Table 8. Accessing High Number of Accounts

Value	Type
Yes	Accessing more accounts
No	Accessing only account

4.2.3.4. Operations Data Table

Operations data contain 8 features plus a target column. Operations data also has a target column labelled as 'Group'. Similar to Login data, there are two labels in this column, 'G' and 'U'. This data has 11105 users. Successfully logged in users from login data table are part of operations data table. Hence the number is much lower than login data table. Genuine users are 8014 while ingenuine users are 3091. The following table shows the features in operations data.

Table 9. Operations Data Features

ID	Login DT	Scanning Imp Ports	Carrying Virus	Unauthorized Access	Unusual Behavior Ops	Atypical Data Loss	Atypical Freq Usage

The second input data to our model is operations data. This data is collected after users have logged into the system. Users exhibit certain behaviors while using the system. The following Table 10 is a snippet of the operations data table.

Table 10. Sample of Operations Data

Group	ID	Login DT	Scanning Imp Ports	Carrying Virus	Unauthorized Access	Unusual Behavior Ops	Atypical Data Loss	Atypical Freq Usage
G	1	1/13/1900	0	0	1	0	1	0
G	2	1/14/1900	0	0	0	0	0	0
G	3	3/18/1900	0	1	1	0	0	0
G	4	3/28/1900	0	0	0	1	0	0
G	5	3/20/1900	0	0	0	0	1	0
G	6	03/06/00	0	0	0	0	0	0
G	7	1/14/1900	1	0	0	1	0	1
G	8	3/19/1900	1	0	0	0	0	0
G	9	1/25/1900	0	1	0	0	0	0

These behaviors are captured to analyze the malicious intentions of the users. Operations data has 7 features which determine the malicious attempts of users. They are:

a. Login ID: Users who have succeeded in logging into system have already been assigned a ID number. Those users are a part of this operations table. Users who could not log in, are not taken in this table.

b. Scanning of Important Ports: When users have been found to scan important ports in the network without required access, those users have malicious intent. This feature is an important feature. The column takes a binary value.

Table 11. Scanning of Important Ports

Value	Type
Yes	Scanning Ports
No	None

c. Carrying Virus: This is also an important feature. When users have been found installing a virus in the system, those users definitely need to have either restricted or no access at all. This column accepts binary value.

Table 12. Carrying Virus

Value	Type
Yes	Carrying Virus
No	None

d. Unauthorized Access: This is yet another important feature that needs to be prioritized. If users are illegally or illegitimately accessing a user account or a file in the system etc. those users need to be quarantined or totally be removed from the system. This column takes binary value as input.

Table 13. Unauthorized Accessing to System

Value	Type
Yes	Unauthorized Access
No	None

e. Unusual Behavior Operations: This feature exhibits users unusual behavior operations. Unusual behavior operations can be a wide range of operations such as accessing the system in unusual times which user has not done previously, usage of a new device previously not recorded in the system, accessing files previously not accessed etc. When any of these behaviors are detected, system administrators need to track these kinds of users. This column accepts binary value.

Table 14. User's Unusual Operations

Value	Type
Yes	Unusual Operations Behavior detected
No	None

f. Atypical Data Loss: This also an important feature. With the detection of users deleting file system which users are not authorized, raises a flag. This column inputs binary value.

Table 15. Performing Unauthorized DML Operations

Value	Type
Yes	Data Loss Observed
No	None

g. Atypical Frequency Usage: This column indicates whether users have accessed the system in unusual times and unusual number of times. These might indicate masquerade attackers. This column accepts binary value.

Table 16. Unusual Frequency Usage

Value	Type
Yes	Atypical Frequency of usage is observed
No	None

4.2.3.5. Processing the Input Data

The provided data is normalized and checked for any null values. Sometimes values in certain features (or columns) are missed. These missed or absent or null values are discarded. Data type conversions are also done in this step.

4.2.3.6. Features Extraction using Boruta algorithm

In order to improve the accuracy of machine learning algorithms and to remove redundancy, feature extraction is performed. The new dataset contains 12 features. These 12 features data are used to train and test four machine learning algorithms. From the combined data, we have 14 features. In order to reduce redundancy and improve machine learning algorithm performance and accuracy, feature extraction is done. This task is accomplished using Boruta algorithm. These above columns from login database and operations database are taken as input to our model. This data is a labelled data. In order to better understand which of these of these columns is more important than others, we have performed feature extraction from both these tables. Feature extraction has been performed using Boruta algorithm.

Our input data consists of high-dimensional data with about 14 features to select from. In order to extract useful information and reduce redundancy among these features, a good statistical method needs to be applied to reduce dimensionality. This extraction is important to train our model better. This is done by inputting only non-correlated and non-redundant data to the model. This way, the model is generalized versus overfitting the training data. This also enhances the accuracy of the model and reduces the noise and improves performance metric.

Other commonly used dimensionality reduction techniques are PCA (Principal Component Analysis), Singular Value decomposition method [53] etc. None of these methods have been chosen for classification, instead Boruta algorithm was chosen for classification. The following are the reasons for this choice:

1. PCA and Singular Value decomposition method are unsupervised learning algorithms. These algorithms do not take into account, the relation between feature values and target values. Instead they take variance in data.
2. These techniques take into account certain assumptions like normality which requires transformations before applying them to our data.

The following diagram explains the feature extraction using Boruta algorithm. The total number of columns given as input are 14 columns. The output features extraction from this model are 12. The input features are denoted by x_1, x_2, x_3, …. x_{14}. A function with the target column(s) needs to be established and this is denoted by y_1, y_2, y_3, …. y_{14}. The observed relationship between features and target column is represented as (x_1, y_1), (x_2, y_2), (x_3, y_3) …. ($x_{14,}$ y_{14}). The extracted features from this process are 12.

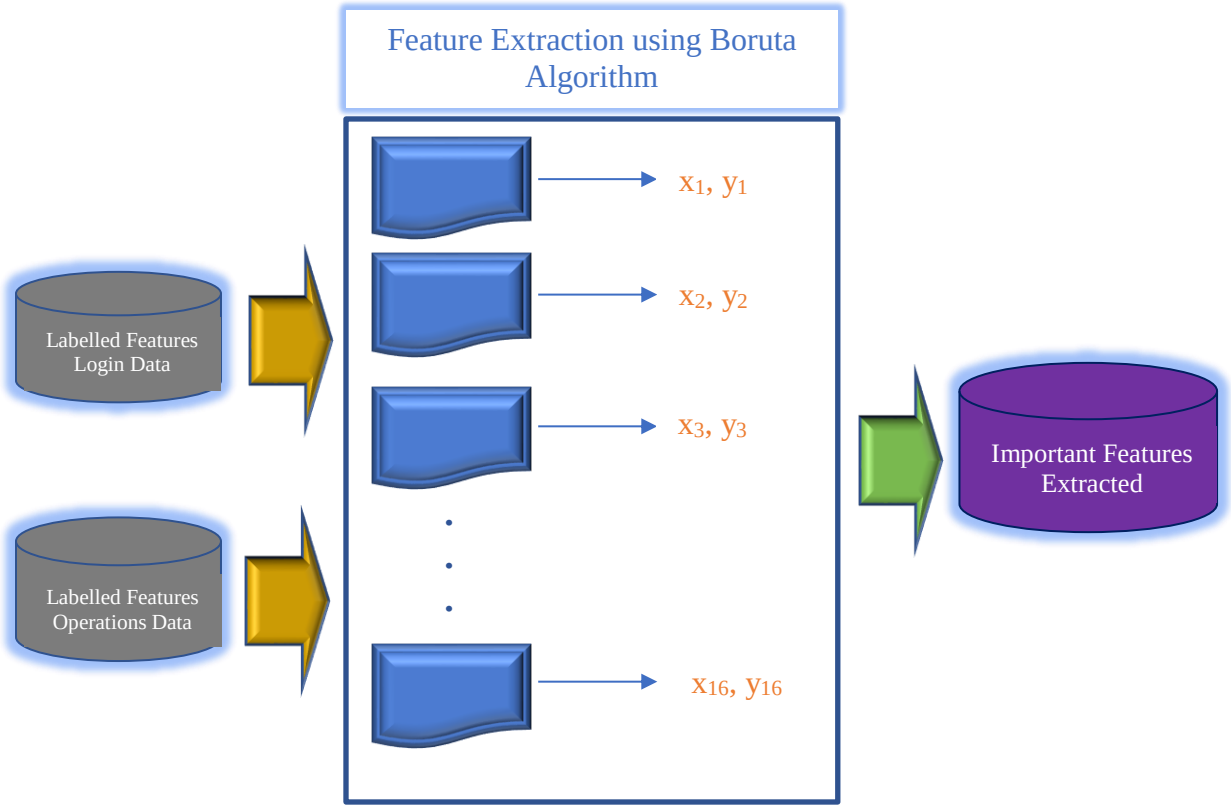

Figure 12. Feature Extraction using Boruta Algorithm

4.2.3.7. Combining the Login and Operations Input Data

Login data table and Operations data table are joined to form as one input. The join is performed on User ID and Login Date. Login data is taken as a join column, as users can login to the system multiple times. To pin down the user using the system for a particular session, date is an important factor to join. Input Data used in the model is split into two kinds. As we could not find relevant data from any repositories, we created simulated data. This data simulation is done using SAS.

4.2.3.8. User Behavior Trust Calculation Prediction Model

The following Figure 13 is the User Behavior Trust Calculation Prediction model.

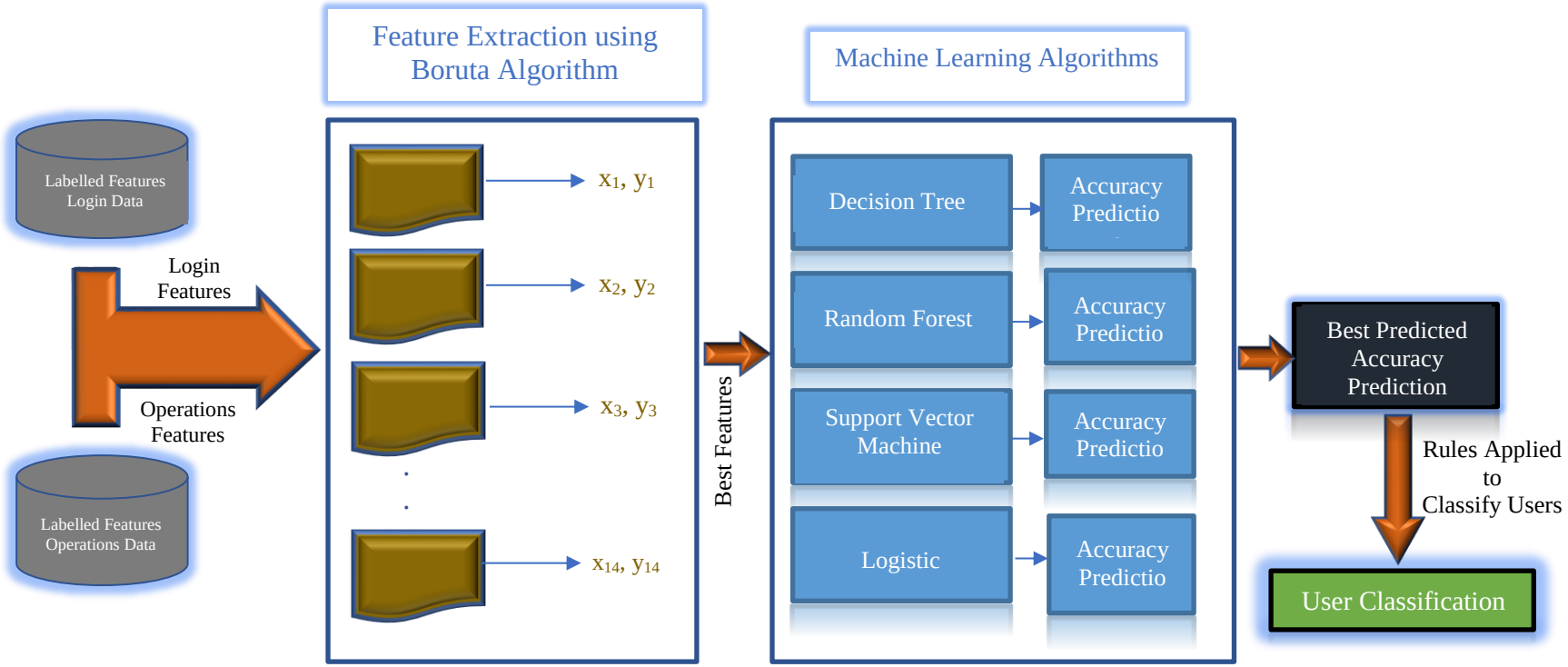

Figure 13. User Behavior Trust Calculation Prediction Model

4.2.3.9. Accuracy Predictions from four Machine Learning Algorithms

Using four machine learning algorithms: Accuracy of the model is predicted using four machine learning algorithms. They are

a. Binomial Logistic Regression

b. Decision Tree

c. Support Vector Machine

d. Random Forest

Each machine learning algorithm is given the extracted features as the input. Each model is trained and tested using k-Fold cross-validation. Accuracy is predicted by each model is stored in a file.

These extracted features form the input data for our model. Final number of extracted features from the data are 12 features. After feature extraction is done, this new database is our input to the model. The second step in the model is to find out the prediction score. After obtaining the important features from the Boruta algorithm, the next step in the model is to obtain the accuracy. This accuracy later is applied to fuzzy logic model to obtain the user's trust category.

For the purpose of predicting accuracy, we have taken four different machine learning algorithms. The following four are the chosen algorithms.

1. Logistic Regression algorithm
2. Decision Tree algorithm
3. Support Vector Machine (SVM) algorithm and
4. Random Forest algorithm

Each machine learning algorithm has been provided the important extracted features as input. To train each supervised machine learning algorithm, it first needs to be trained and then

tested. Algorithms such as Random Sampling are used to split the data randomly. This random split can happen anywhere. Generally, the data is split into 80:20 ratio. There are few problems with this approach.

1. The training Data set is manually split into 80:20. 80 for training and 20 for testing the data. The training and test are only performed once. The presence or absence of a single outlier can greatly affect the mean square error. As a result, machine learning algorithm might perform with a high accuracy for training data and can fail terribly with general data or test data.
2. To maximize the accuracy rate, the split training data and test data needs to be maximized. To get better training results, the training data set needs to have maximum data items. To have the best validation, test data set needs to be maximized. The dilemma is what ratio needs to be used.

A better approach is to use k-Fold cross validation, which gives more precise estimate of the true out-of-sample error. This cross-validation algorithm shuffles the data randomly and splits into k-groups. A group is hold-out and the remaining groups are used for training the data. This is used for training the model and the hold-out group to validate the machine leaning algorithm. This way we have k training sets and k test sets. This implies that the training data set, and the test data are of the same size.

The k-Fold approach is applied on all the four machine learning algorithms. The k value is taken as ten. This means we have 10 training data sets and 10 test data sets. All the four machine learning algorithms are run 10 -Fold. The accuracy obtained from each algorithm is later compared to choose the best algorithm for the corresponding data set. The following Figure 14 outlines the Logistic Regression model.

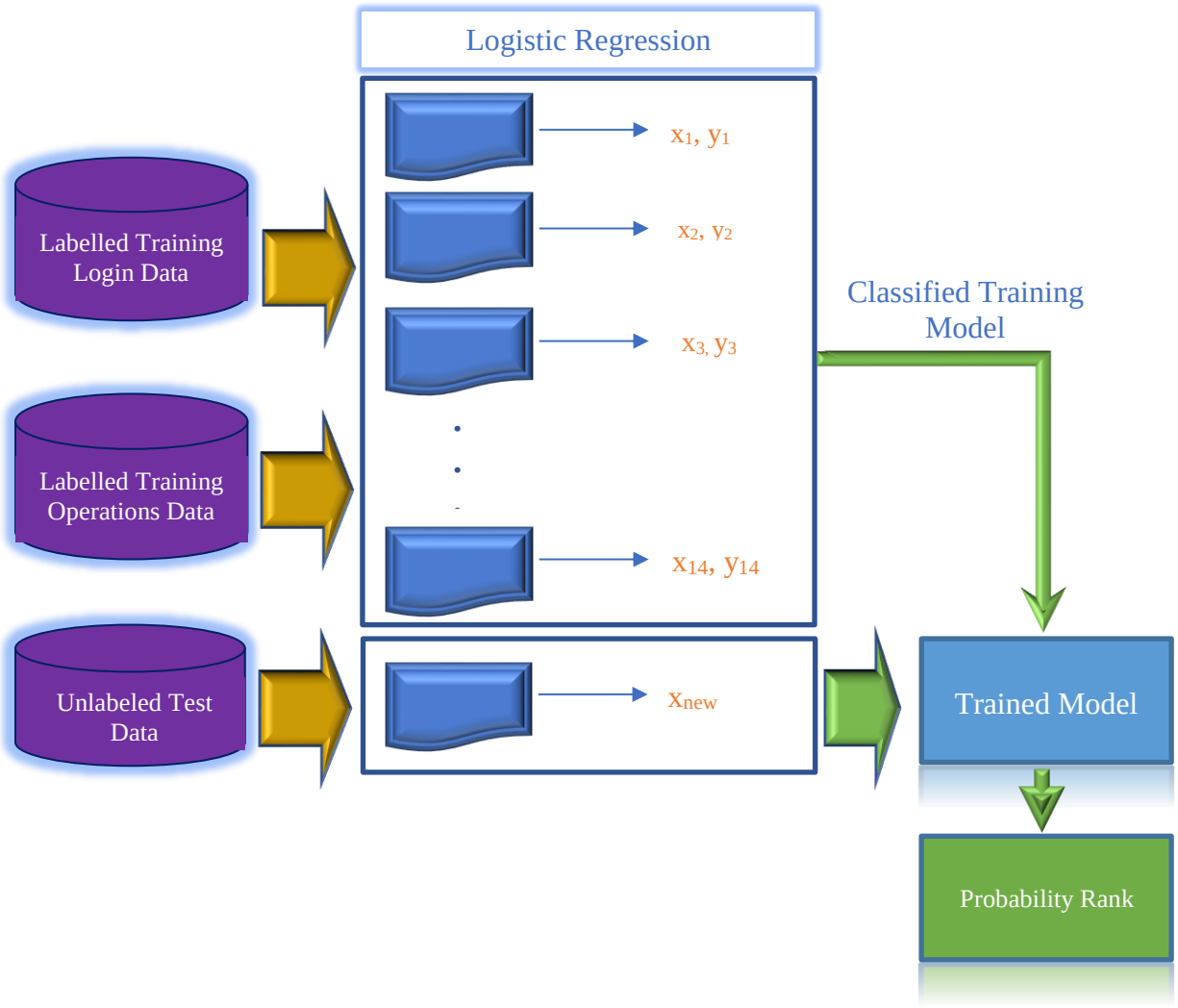

Figure 14. Logistic Regression Classifier

The following Figure 15 describes in detail about the working of Decision Tree algorithm. The input parameters for training the model are the labelled login data and operations data. The test data is unlabeled data that comprises of both login and operations data. Each feature is labelled $x_1, x_2, x_3, \ldots x_{12}$. A function is established for each feature with the target column. After the model has been trained with the labelled data, it is tested for accuracy with the unlabeled data.

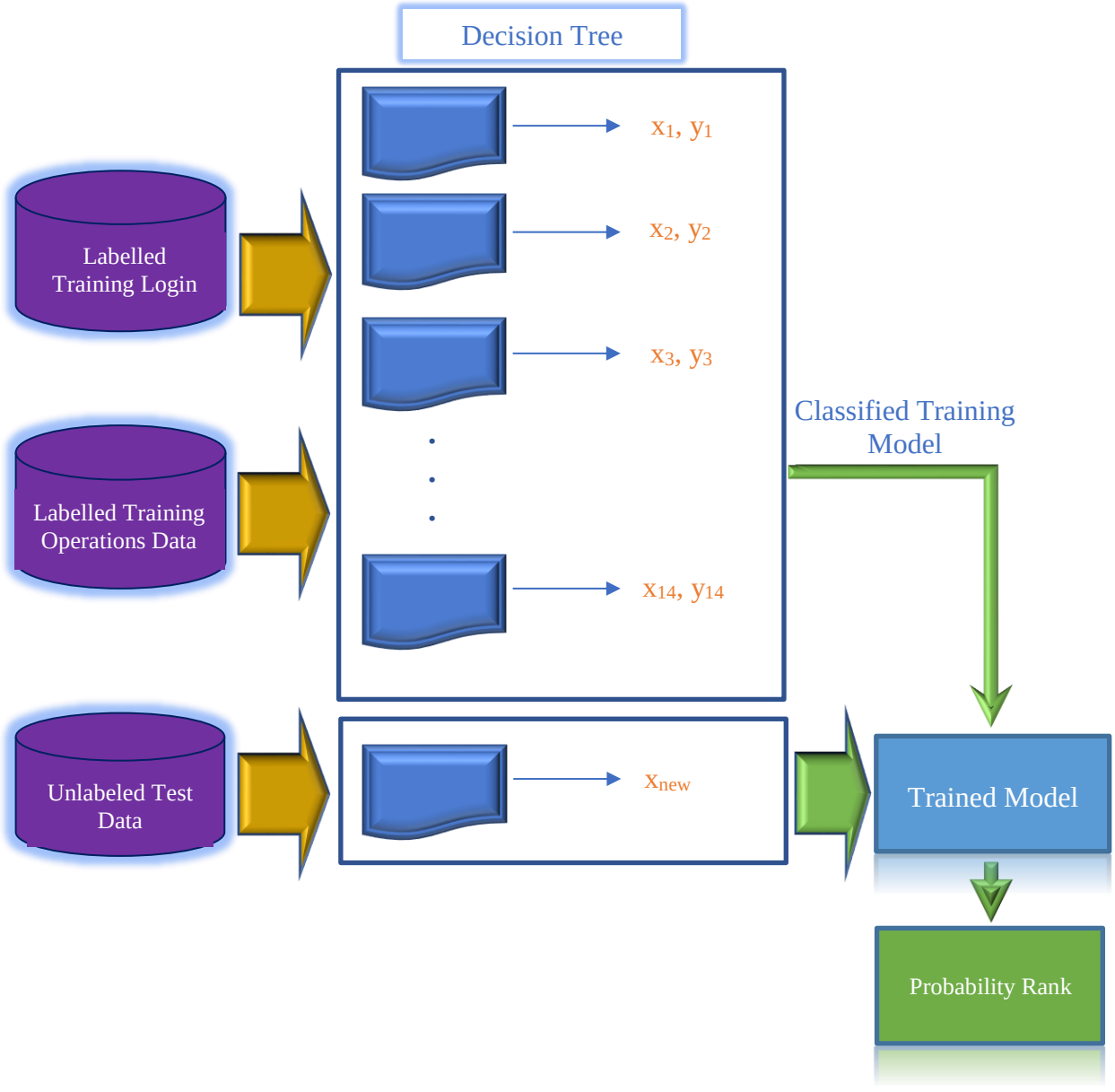

Figure 15. Decision Tree Model

The following Figure 16 describes in detail about the working of Support Vector Machine algorithm. The accuracy from SVM is 97.92 for k-fold equals to 3.

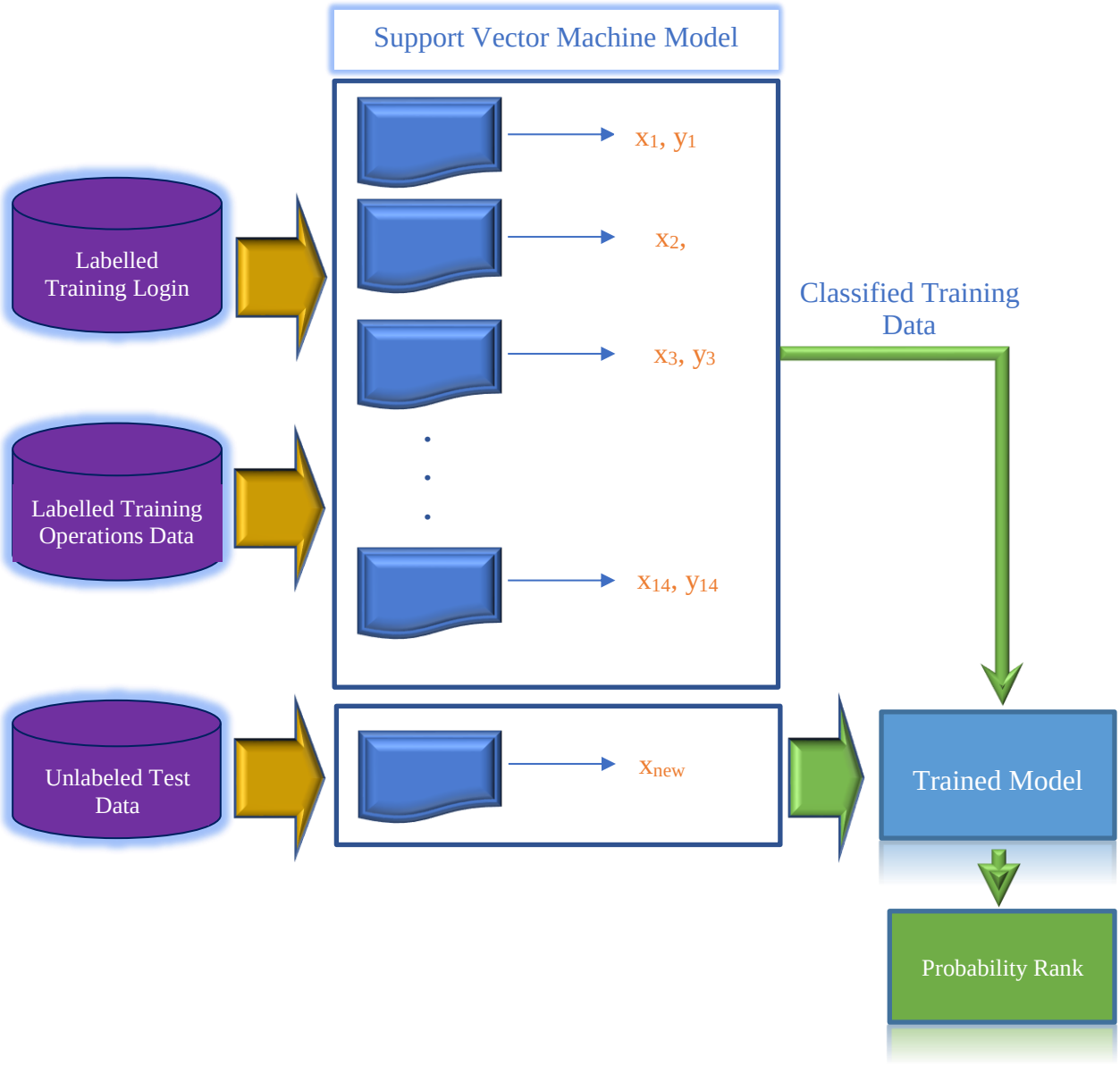

Figure 16. Support Vector Machine Model

The following Figure 17 describes in detail about the working of Random Forest Machine algorithm. The accuracy for Random Forest for k-fold equals 3 is 97.65.

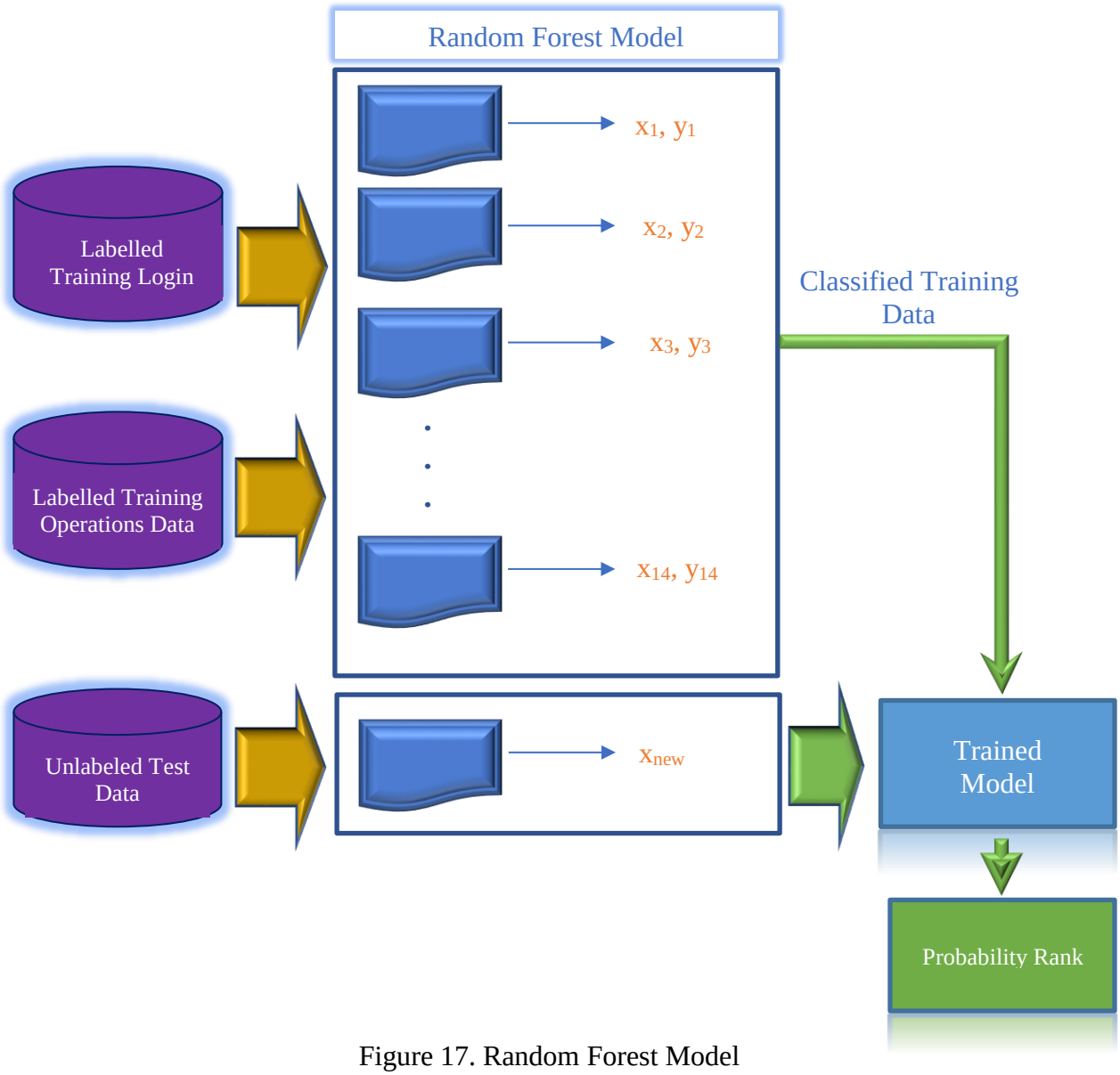

Figure 17. Random Forest Model

4.2.3.10. Choosing the Best Predicted Accuracy

The accuracy obtained from each machine learning algorithm is stored in a file. The highest accuracy predicted, and the corresponding algorithm is taken for user classification. This accuracy is used to classify the users in six different categories. They are

a. Very High Trust
b. High Trust
c. High Medium Trust
d. Low Medium Trust
e. Low Trust
f. Very Low Trust

The output from each four of the machine learning algorithms is accuracy prediction. Accuracy is a metric which evaluates classification models, in our case, machine learning algorithms. Informally, accuracy prediction gives us the degree of closeness to the actual or true value.

$$\text{Accuracy} = \frac{\text{Number of Correct predictions}}{\text{Total number of predictions}}$$

This is accomplished by training the model followed by testing the model using k-Fold cross validation. Results (accuracy prediction) obtained from each machine learning algorithm are different from one another. This accuracy can change depending on the provided data, this means that the input data can affect the accuracy of machine learning method. One model might perform better than others. It is hard to predict which algorithm will function better than others. The question then arises is which machine learning algorithm out of the four is chosen to maximize the prediction.

To solve this problem, we have automatized the selection of algorithm. The model that predicts highest accuracy amongst four machine learning algorithms is automatically selected. We do not have to manually review accuracy prediction of each algorithm to make a final selection. Using the k-Fold cross-validation method, performance of each of the four machine learning algorithms is checked per k-Fold of data. The algorithm that has the highest accuracy is selected to proceed further in our model. In our model, Logistic Regression model has the highest accuracy prediction score amongst all the four models. This accuracy prediction scores forms the input to our fuzzy logic model.

4.2.3.11. Classifying Users into Six Categories

The prediction scores from the machine learning models forms the bases for classifying the users. There are 6 buckets that users are classified into. The scoring is done based on the following rules:

Table 17. User Classification Categories

Value	**Category**
0.9-1.0	Very High Trust
0.75-0.9	High Trust
0.5-0.75	High Medium Trust
0.25-0.5	Low Medium Trust
0.1-0.25	Low Trust
0.0-0.1	Very Low Trust

Prediction scores of users categorize users into the buckets as mentioned above. Users with value of 0.9 or higher are considered a ‘Very Hight Trust’ users, users with value of 0.3 are placed in ‘Low Medium Trust’ buckets etc. These values can be changed by the service providers according to their requirement. Users will fall into the buckets according to the new values provided by the service providers.

Placing the users in these 6 buckets is the final output of our model. According to these output categories, service providers can take the necessary actions.

5. RESULTS

Service providers such as cloud service providers, storage service providers, email service providers etc. when providing services to users, also to need to focus on security of users and resources as well. Security can be enhanced when the user behavior is added into the existing security technologies. The main motivation behind designing and building this model, is to evaluate the trustworthiness of users who are utilizing services. If user behavior can be detected during early stages, an active attack can be immediately avoided.

The final output from our model is the six different classes of trustworthiness of users. Each user from the input data falls into these six categories. The following are 6 categories based on trust value.

Table 18. User Behavior Trust Categories

Trust Value	**Category**
0.9-1.0	Very High Trust
0.75-0.9	High Trust
0.5-0.75	High Medium Trust
0.25-0.5	Low Medium Trust
0.1-0.25	Low Trust
0.0-0.1	Very Low Trust

Four machine learning algorithms have been applied in our model, out of which one model with best accuracy has been chosen. From our experiment, Logistic Regression model stands on the top of all models. The accuracy prediction is 98.109.

Table 19. Predicted Accuracy of each Machine Learning Algorithm

Logistic Regression	**Decision Tree**	**SVM**	**Random Forest**	**Max Accuracy**
98.10981	94.23942	97.92979	97.65977	**98.10981**

Logistic Regression has scored the highest accuracy amongst all the four machine learning algorithms. The columns '*Log Reg Prob Rank*' implies the Logistic Regression Probability Rank. It indicates percentage of accuracy when compared to true or actual value. If the value is 0.97,

this value indicates the degree of closeness to the actual value. The column '*Log Reg Prediction*' implies Logistic Regression Prediction. This column indicates whether the output is zero or 1. It indicates whether the calculated value is zero or one. The value zero indicates that the user is not a genuine or good user, whereas the value one indicates that the user is genuine. The column '*Log Prediction Check*' compares the actual value with the predicted value. If the predicted value matches the actual value, the log prediction check is true, else false. For instance, if the logistic regression model predicted a zero but the actual value is one, then the logistic prediction check value is false. The final column is 'User Classification', where users are placed in the corresponding buckets based on the trust value.

The following Table 20. Logistic Regression Prediction Model Table 20 is a snippet of the output table for Logistic Regression model.

Table 20. Logistic Regression Prediction Model

IP	DEVICE	LOGIN SUCCESS	USER LOGIN ATTEMPTS	USER PROXY	HIGH ACCOUNTS	SCANNING IMP PORTS	CARRYING VIRUS	UNAUTH ACCESS	UNUSUAL BEHAVIOR	DATA LOSS	FREQ USAGE	USER GROUP	LOG REG PROB RANK	LOG REG PREDICTION	LOG REG PREDICTION CHECK	USER CLASSIFICATION
1	2	1	1	0	0	0	1	1	0	0	0	1	1	1	TRUE	Very High Trust
1	4	1	1	0	1	0	1	0	0	0	0	1	1	1	TRUE	Very High Trust
1	3	1	1	0	0	0	0	1	0	0	0	1	1	1	TRUE	Very High Trust
1	2	1	1	0	0	0	0	0	0	0	0	1	1	1	TRUE	Very High Trust
1	4	1	2	0	0	0	0	1	0	0	1	1	0.97	1	TRUE	Very High Trust
0	4	1	1	0	1	0	0	0	0	0	1	1	0.98	1	TRUE	Very High Trust
1	4	1	1	0	1	1	0	0	0	0	0	1	1	1	TRUE	Very High Trust
1	4	1	1	0	0	0	1	1	0	0	1	1	0.99	1	TRUE	Very High Trust
1	3	1	1	0	0	0	0	0	0	0	0	1	1	1	TRUE	Very High Trust
1	2	1	2	0	1	0	0	0	0	1	0	1	0.99	1	TRUE	Very High Trust
1	4	1	1	1	0	1	1	1	0	1	1	0	0.13	0	TRUE	Low Trust

The following is the console output of the Logistic Regression model.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.9606  -0.0096   0.0282   0.0812   5.3279

Coefficients:
                   Estimate Std. Error z value             Pr(>|z|)
(Intercept)          7.4946     0.7759   9.660 < 0.0000000000000002 ***
IP1                  2.0704     0.1433  14.449 < 0.0000000000000002 ***
DEVICE2              1.6359     0.3316   4.933          0.000000809 ***
DEVICE3             -0.2360     0.3288  -0.718             0.472907
DEVICE4              0.2926     0.2740   1.068             0.285627
LOGIN_SUCCESS1       2.5453     0.6653   3.825             0.000131 ***
USER_LOGIN_ATTEMPTS -2.5120     0.1106 -22.709 < 0.0000000000000002 ***
USER_PROXY1         -2.3087     0.1466 -15.746 < 0.0000000000000002 ***
HIGH_ACCOUNTS1      -2.0638     0.1434 -14.396 < 0.0000000000000002 ***
SCANNING_IMP_PORTS1 -2.0574     0.1421 -14.482 < 0.0000000000000002 ***
CARRYING_VIRUS1     -1.5977     0.1406 -11.364 < 0.0000000000000002 ***
UNAUTH_ACCESS1      -1.9153     0.1429 -13.401 < 0.0000000000000002 ***
UNUSUAL_BEHAVIOUR1  -2.0480     0.1426 -14.361 < 0.0000000000000002 ***
DATA_LOSS1          -1.8681     0.1426 -13.104 < 0.0000000000000002 ***
FREQ_USAGE1         -2.0593     0.1444 -14.260 < 0.0000000000000002 ***

Null deviance: 11833.0 on 9991 degrees of freedom
Residual deviance:  1527.8 on 9977 degrees of freedom
AIC: 1557.8

Number of Fisher Scoring iterations: 8
```

Figure 18. Console Output of the Logistic Regression Model

The columns ‘Decision Tree Prob rank’ implies the Decision Tree probability. It indicates accuracy percentage as actual value. The column ‘Decision Tree Prediction’ indicates whether the output is zero or 1. The value zero indicates that the user is not a genuine or good user, whereas the value one indicates that the user is genuine. The column ‘Decision Tree Prediction Check’ compares the actual value with the predicted value. If the predicted matches the actual value, the log prediction check is true, else false.

Table 21. Decision Tree Prediction Model Output

IP	DEVICE	LOGIN SUCCESS	USER LOGIN ATTEMPTS	USER PROXY	HIGH ACCOUNTS	SCANNING IMP PORTS	CARRYING VIRUS	UNAUTH ACCESS	UNUSUAL BEHAVIOR	DATA LOSS	FREQ USAGE	USER GROUP	DECISION TREE PROB RANK	DECISION TREE PREDICTION	DECISION TREE PREDICTION CHECK
1	2	1	1	0	0	0	1	1	0	0	0	1	0.97	1	TRUE
1	4	1	1	0	1	0	1	0	0	0	0	1	0.97	1	TRUE
1	3	1	1	0	0	0	0	1	0	0	0	1	0.97	1	TRUE
1	2	1	1	0	0	0	0	0	0	0	0	1	0.97	1	TRUE
1	4	1	2	0	0	0	0	1	0	0	1	1	0.04	0	FALSE
0	4	1	1	0	1	0	0	0	0	0	1	1	0.97	1	TRUE
1	4	1	1	0	1	1	0	0	0	0	0	1	0.97	1	TRUE
1	4	1	1	0	0	0	1	1	0	0	1	1	0.97	1	TRUE
1	3	1	1	0	0	0	0	0	0	0	0	1	0.97	1	TRUE
1	2	1	2	0	1	0	0	0	0	1	0	1	0.78	1	TRUE

Table 22. SVM Prediction Model Output

IP	DEVICE	LOGIN SUCCESS	USER LOGIN ATTEMPTS	USER PROXY	HIGH ACCOUNTS	SCANNING IMP PORTS	CARRYING VIRUS	UNAUTH ACCESS	UNUSUAL BEHAVIOR	DATA LOSS	FREQ USAGE	USER GROUP	SVM PROB RANK	SVM PREDICTION	SVM PREDICTION CHECK
1	2	1	1	0	0	0	1	1	0	0	0	1	1	1	TRUE
1	4	1	1	0	1	0	1	0	0	0	0	1	1	1	TRUE
1	3	1	1	0	0	0	0	1	0	0	0	1	1	1	TRUE
1	2	1	1	0	0	0	0	0	0	0	0	1	1	1	TRUE
1	4	1	2	0	0	0	0	1	0	0	1	1	0.97	1	TRUE
0	4	1	1	0	1	0	0	0	0	0	1	1	0.98	1	TRUE
1	4	1	1	0	1	1	0	0	0	0	0	1	1	1	TRUE
1	4	1	1	0	0	0	1	1	0	0	1	1	0.98	1	TRUE
1	3	1	1	0	0	0	0	0	0	0	0	1	1	1	TRUE
1	2	1	2	0	1	0	0	0	0	1	0	1	0.99	1	TRUE

The column ‘SVM Prob Rank’ is the percentage of accuracy as compared to the actual value. The column ‘SVM Prediction Rank’ indicates whether the calculated value is zero or one. Similar to other previous models, the value zero indicates that the user is not a genuine or good

user, whereas the value one indicates that the user is genuine. The column ‘SVM Prediction Check’ compares the actual value with the predicted value.

The column ‘Random Forest Prediction Rank’ is the percentage of accuracy as compared to the actual value. The column ‘Random Forest Prediction Rank’ indicates whether the calculated value is zero or one. The column ‘Random Prediction Check’ gives output as true or false.

Table 23. Random Forest Prediction Model

IP	DEVICE	LOGIN SUCCESS	USER LOGIN ATTEMPTS	USER PROXY	HIGH ACCOUNTS	SCANNING IMP PORTS	CARRYING VIRUS	UNAUTH ACCESS	UNUSUAL BEHAVIOR	DATA LOSS	FREQ USAGE	USER GROUP	RF PROB RANK	RF PREDICTION	RF PREDICTION CHECK
1	2	1	1	0	0	0	1	1	0	0	0	1	1	1	TRUE
1	4	1	1	0	1	0	1	0	0	0	0	1	1	1	TRUE
1	3	1	1	0	0	0	0	1	0	0	0	1	1	1	TRUE
1	2	1	1	0	0	0	0	0	0	0	0	1	1	1	TRUE
1	4	1	2	0	0	0	0	1	0	0	1	1	0.93	1	TRUE
0	4	1	1	0	1	0	0	0	0	0	1	1	1	1	TRUE
1	4	1	1	0	1	1	0	0	0	0	0	1	1	1	TRUE
1	4	1	1	0	0	0	1	1	0	0	1	1	1	1	TRUE
1	3	1	1	0	0	0	0	0	0	0	0	1	1	1	TRUE
1	2	1	2	0	1	0	0	0	0	1	0	1	0.89	1	TRUE

The following bar chart indicates the percentage of accuracy for each machine learning algorithm. This accuracy is taken when the k-fold is 3.

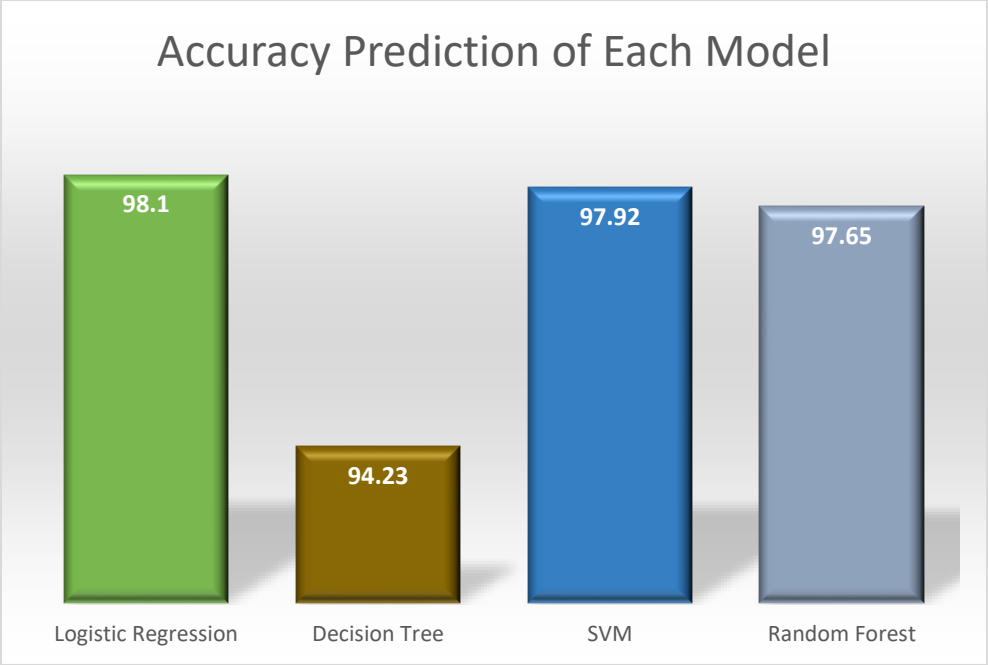

Figure 19. Accuracy Prediction of four Machine Learning Algorithms

5.1. Model Comparison

We compared our model to two similar models. One of them is Neural Network based model and another is ILSTM.

Improved Long Short-Term Memory model [37] is an supervised learning model proposed by Nathezhtha et all. This model can classify the user behavior into normal and abnormal categories. This model was able to achieve a very good accuracy of 90%.

Neural Network based model [38] as proposed by Martin et all is a supervised learning algorithm. The goal of the work is to detect suspicious user behavior from the data. This model attained an impressive accuracy of 98%.

Our model though takes a supervised learning approach has achieved an accuracy of 98.1%. This suggests that our model can perform well in categorize users as genuine or ingenuine and also calculate trust value of each user.

Table 24. Model Comparison between ILSTM and our Model

Methods Used	Accuracy
ILSTM[2]	90%
Our model	98.1%

Table 25. Model Comparison between Neural Network based approach and our Model

Methods Used	Accuracy
Neural Network based Model[1]	98%
Our model	98.1%

1. Identifying Suspicious User Behavior using Neural Network proposed by Martin et al
2. ILSTM - Improvised Long Short-Term Memory Model proposed by Nathenzhtha et al

The following diagram is the final output of our model. Given input data, we have classified the users into six different categories. From the following graph, we can say that majority of users fell into Very High Trust category, followed by Very Low Trust. This output is in accordance with our input. The combined data (login and operations) had about 75% genuine users and 25% ingenuine users. As a result, the final output was the reflection of the input data.

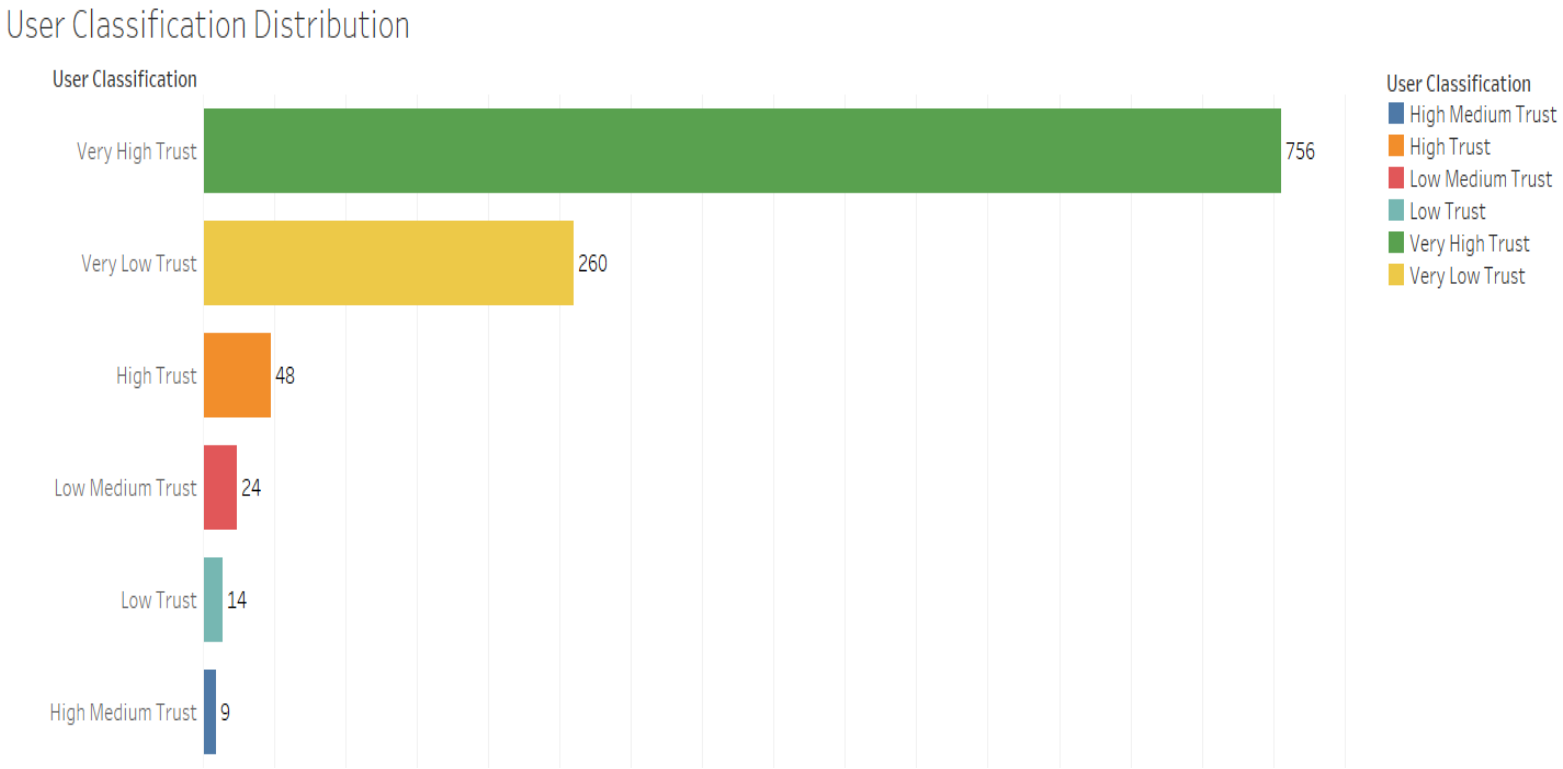

Figure 20. User Classification into six Categories

6. CONCLUSION

The growth of modern communication mediums has not only facilitated the growth of communication but also brought in the security challenges. Pervasive computing, cloud computing and traditional computers still use conventional access methods to validate the users. This in the past has proven to be vulnerable, as masquerade attacks and insider attacks are immune to security mechanisms.

Our model has incorporated user behavior trust aspect in validating the users. It also triggers an alarm for insider attacks. Any anomalies detected with respect to user behavior, is a good candidate user for malicious users.

After extracting features from the input data, the data was fed to four different machine learning algorithms to determine the accuracy of each machine learning algorithm. We find out the best performing model from the accuracies obtained from the four model. The corresponding model is chosen for further to test the data. Logistic Regression model accuracy was 98. Inspired from fuzzy logic rules, we have written rules to categorize the users into six different buckets.

The model can be tailored to the requirements of service providers. Machine learnings can be changed or the number of machine learning algorithms can be either increased or decreased according to the input data. Some input data require certain machine learning algorithms. This can be tailored according to the data. The fuzzy rules can also be changed as needed.

In conclusion, this dissertation is a significant research work that proposes new model for cloud and pervasive computing environments.

REFERENCES

[1] H. T. Company, "5G: A technology vision," Huawei Technologies Co.," 2013. [Online]. Available: https://www.huawei.com/ilink/en/download/HW_314849. [Accessed 15 April 2019].

[2] S. Electronics, "5G Vision," February 2015. [Online]. Available: https://www.samsung.com/global/business/networks/insights/white-paper/5g-vision/. [Accessed 15 April 2019].

[3] N. Networks, "Looking Ahead to 5G," 25 June 2014. [Online]. Available: https://eucnc.eu/files/keynotes/Moiin.pdf. [Accessed 15 April 2019].

[4] A. Kumar, Y. Liu, J. Sengupta and Divya, "Evolution of Mobile Wireless Communication Networks: 1G to 4G," *International Journal of Electronics & Communication Technology,* vol. 1, no. 1, pp. 68-72, 2010.

[5] M. Rouse, "IoT Agenda," Tech Target, 8 November 2016. [Online]. Available: https://internetofthingsagenda.techtarget.com/definition/pervasive-computing-ubiquitous-computing. [Accessed 8 November 2018].

[6] K. Henricksen, J. Indulska and A. Rakotonirainy, "Modeling Context Information in Pervasive Computing Systems," *International Conference on Pervasive Computing,* pp. 167-180, 2002.

[7] G. V. Hulme, The Threat from Inside, Information Week, 2002.

[8] E. D. Shaw, K. G. Ruby and J. M. Post, "The insider threat to Informaiton Systems," *Semantic Scholar,* 1998.

[9] M. Blaze, J. Feigenbaum, J. Ionnidis and A. D. Keromytis, "The role of trust managment in distributed systems security," *Security issues in mobile and distributed systems,* 1999.

[10] C. E. Tsirakis, P. Matzoros, P. Sioutis and G. S. Agapiou, "Load balancing in 5G Networks," *MATEC Web of Conferences,* vol. 125, p. 6, 2017.

[11] C. Askarian and H. Beigy, "A Survey for Load Balancing in Mobile WiMAX Networks," *Advanced Computing International Journal (ACIJ),* vol. 3, no. 2, pp. 119-137, March 2012.

[12] S. Furman, "Building Trust," usability.gov, 1 Sep 2018. [Online]. Available: https://www.usability.gov/get-involved/blog/2009/09/building-trust.html. [Accessed 15 April 2019].

[13] Y. D. Wang and H. H. Emurian, "An overview of online trust: Concepts, elements, and implications," *Science Direct,* vol. 21, no. 1, pp. 105-125, 2005.

[14] P. Andras, L. Esterle, M. Guckert, T. A. Han, P. R. Lewis, K. Milanovic, T. Payne, C. Perret, J. Pitt, S. T. Powers, N. Urquhart and S. Wells, "Trusting Intelligent Machines," *IEEE Technology and Society,* vol. 37, no. 4, pp. 76-83, 2018.

[15] P. Paganini, "Cost of cybercrime will grow from $3 trillion (2015) to $6 trillion by 2021Security Affairs," Security Affairs, 28 August 2016. [Online]. Available: http://securityaffairs.co/wordpress/50680/cyber-crime/global-cost-of-cybercrime.html. [Accessed 12 December 2018].

[16] K. E. Nygard, M. M. Chowdhury, K. Kambhampaty and P. Kotala, "Cybersecurity Materials for K-12 Education," *Midwest Instruction and Computing Symposium,* 2018.

[17] K. Kambhampaty, M. Alruwaythi, M. M. Chowdhury and K. E. Nygard, "Trust and its Influence on Technology," in *Midwest Instruction and Computing Symposium*, 2019.

[18] F. Belanger, "College of Business," Virgina Tech, 2011. [Online]. Available: https://www.magazine.pamplin.vt.edu/fall11/passwordsecurity.html. [Accessed 20 March 2019].

[19] A. Sasse and I. Flechais, "Usable Security: Why Do We Need It? How Do We Get It?," *Security and Usability,* pp. 13-30, 2005.

[20] S. L. Pfleeger and D. D. Caputo, "Leveraging Behavioral Science to Mitigate Cyber Security Risk," *Science Direct,* vol. 31, no. 4, pp. 597-611, June 2012.

[21] H. M. Knight, Trust in Information Technology, Malden M.A., 2005, pp. 329-331.

[22] M. Alruwaythi, K. Kambhampaty and K. E. Nygrd, "User Behavior Trust Modeling in Cloud Security," *IEEE Computational Science and Engineering,* 2019.

[23] A. Team, "Amazon Web Services," Amazon, 20 july 2008. [Online]. Available: https://status.aws.amazon.com/s3-20080720.html. [Accessed 2019 April 25].

[24] K. Mahaffey, "Techcrunch," Tech Crunch, 2010. [Online]. Available: https://techcrunch.com/2010/06/15/ipad-breach-personal-data/. [Accessed 25 April 2019].

[25] F. Landman, "ReadWrite," Read Write, 21 December 2018. [Online]. Available: https://readwrite.com/2018/12/21/why-the-public-still-doesnt-fully-trust-cloud-computing-and-what-we-can-do/. [Accessed 03 April 2019].

[26] G. Hurlburt, "How Much to Trust Artificial Intelligence," *IEEE Computer Society,* vol. 19, no. 4, pp. 7-11, 2017.

[27] W. Pieters, "Explanation and trust: what to tell the user in security and AI?," *Springer,* vol. 13, no. 1, pp. 53-64, 2010.

[28] P. E. Johnson, S. Grazioli and K. Jamal, "Fraud detection: Intentionality and deception in cognition," *Accounting, Organization and Society,* vol. 18, no. 5, pp. 467-488, 1993.

[29] K. Kambhampaty, M. Alruwaythi, M. M. Chowdhury and K. E. Nygard, "Identifying Malicious Users Through Behavior," in *Midwest Instruction and Computing Symposium,* Fargo, ND, 2019.

[30] L. Wen, P. Lingdi, L. Kuijun and C. Xiaoping, "Trust Model of Users' Behavior in Trustworthy Internet," *WASE International Conference on Information Engineering,* 2009.

[31] Z. Liu and D. Peng, "User Behavior Identification for trust Management in Pervasive Computing Systems," *IEEE,* pp. 65-72, 2007.

[32] M. Alruwaythi, K. Kambhampaty and K. E. Nygard, "User Behavior and Trust Evaluation in Cloud Computing," *Proceedings of 34th International Confer-,* vol. 58, pp. 378-386, 2019.

[33] J. Ma and Y. Zhang, "Research on Trusted Evaluation Method of User Behavior Based on AHP Algorithm," *International Conference on Information Technology in Medicine and Education,* 2015.

[34] R. Yang and X. Yu, "Research on Way of Evaluating Cloud End User Behaviors Credibility Based on the Methodology of Multilevel Fuzzy Comprehensive Evaluation," *International Conference on Software and Computer Applications,* pp. 165-170, 2017.

[35] M. Ali and M. Homayun, "A trust model between cloud entities using fuzzy mathematics," *Journal of Intelligent & Fuzzy Systems,* pp. 1795-1803, 2015.

[36] N. Yang, H. Barringer and N. Zhang, "A Purpose-Based Access Control Model," *Third International Symposium on Information Assurance and Security,* pp. 51-58, 2007.

[37] N. T and V. V, "Cloud Insider Attack Detection Using Machine Learning," *IEEE,* 2018.

[38] M. Ussath, D. Jaeger, F. Cheng and C. Meinel, "Identifying Suspicious User Behavior with Neural Networks," *IEEE 4th International Conference on Cyber Security and Cloud Computing,* 2017.

[39] C. Liu, M. Fan and G. Wang, "Unsupervised behavior evaluation method in trustworthy network," *International Workshop on Education Technology and Computer Science,* 2010.

[40] S. Mavoungou, G. Kaddoum, M. Taha and G. Matar, "Survey on Threats and Attacks on Mobile Networks," *IEEE Access,* vol. 4, pp. 4543 - 4572, July 2016.

[41] A. Chuvakin and E. Heid, "Understanding Insider Threats," Gartner Research, 02 May 2016. [Online]. Available: https://www.gartner.com/en/documents/3303117. [Accessed 15 May 2019].

[42] A. Green, "User Behavior Analytics," Varonis, [Online]. Available: https://www.varonis.com/blog/what-is-user-behavior-analytics/. [Accessed 10 May 2019].

[43] "Machine Learning," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Machine_learning. [Accessed 06 05 2019].

[44] "Medium," Daffodil Software, 30 July 2017. [Online]. Available: https://medium.com/app-affairs/9-applications-of-machine-learning-from-day-to-day-life-112a47a429d0. [Accessed 06 05 2019].

[45] P. Langley and H. A. Simon, "Applications of Machine Learning and Rule Induction," *Institute for the Study of Learning and Expertise,* vol. 38, no. 11, pp. 54-64, Feb 15 1995.

[46] "Decision Tree," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Decision_tree. [Accessed 07 05 2019].

[47] C.-W. Hsu, C.-C. Chang and C.-J. Lin, "A Practical Guide to Support Vector Classification," April 15 2010.

[48] V. Kecman, Support Vector Machines - An Introduction, Berlin, Heidelberg: www.springerlink.com, 2005.

[49] "Logistic Regression," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Logistic_regression. [Accessed 08 05 2019].

[50] T. K. Ho, "Random Decision Forest," *Proceedings of the 3rd International Conference on Document Analysis and Recognition,* vol. 1, pp. 278-282, 1995.

[51] X. Shan and S.-b. Zhang, "Trust assessment approach about user's behaior in complex network environment," *Software Journal,* vol. 33, pp. 64-68, 2012.

[52] "Public Service Announcement," Federal Bureau of Investigation, 02 August 2018. [Online]. Available: https://www.ic3.gov/media/2018/180802.aspx. [Accessed 15 May 2019].

[53] M. Pathak, "Feature Selection in R with the Boruta R Package," Data Camp, 7 March 2018. [Online]. Available: https://www.datacamp.com/community/tutorials/feature-selection-R-boruta. [Accessed 17 05 2019].