Sudan University of Science & Technology

Faculty of Computer Science and Information Technology

Risk Assessment Model for Cloud Computing using Machine Learning Techniques

نموزج تقييم المخاطر للحوسبه السحابيه باستخدام تقنيات تعلم الاله

This Thesis is Submitted in Partial Fulfilment of The Requirements For The Degree of Doctor of Philosophy in Computer Science at Sudan University of Science & Technology

By
Nada Ahmed Mohammednour Eisa

Supervisor
Professor Dr. Ajith AbrahamJune

2016

Dedication

I would like to dedicate this work to my beloved parents, Ahmed and Rogia. I would not reach this level of education without your caring support and influential help. For your sacrifices for me the whole time, I dedicate this work for you. I would like to dedicate this work to my brother and sisters who keep encouraging me and always be happy about my achievements more than me. To my small family my husband Omar, my son Musab, and my daughter Tibyan.

Acknowledgment

I wish to acknowledge the support and help of several people who have been instrumental in some way or the other for making this dissertation. First and foremost, I would like to thank my parents for their never ending support and encouragement throughout my education. My mother has always been my biggest motivation in times of high and low. My father has been an invaluable source of inspiration for me always.

I would like to thank my supervisor Professor Ajith Abraham for his patience, guidance and constant encouragement. It has been a great privilege to work under him. Many thanks to Prof. Izzaldin Mohammed Osman, who be as our Spiritual Father. I thank Mr. Varun Ojha for his help in running some of the experiments required for this thesis.

Special and many thanks must also go to my husband for encouraging and supporting me all these years. My thanks go to all my friends and colleagues for being there for me and understanding my up and down moods.

ABSTRACT

Cloud computing has recently emerged as a new paradigm for hosting and delivering services to the customers over the Internet. A cloud computing system is a set of resources designed to be allocated ad hoc to run applications, rather than be assigned a static set of applications as is the case in client/server computing. Cloud Computing is being introduced and marketed with many attractive promises that are enticing to many companies and managers around the world, such as reduced costs, and relief from managing complex IT infrastructure. Virtualization technologies enable the abstraction and pooling of resources to be shared across the organization, data centers are designed around virtual machines, which are the new atomic units of computing.

Traditionally, it is believed that any connectivity to systems or organizations outside of an organization provides an opening for unauthorized entities to gain access or tamper with information resources. Cloud computing moves computing and data away from desktop and portable PC's into large data center distributed around the world. As a result, this will create a need for a considerable risk assessment approach to manage the various types of risks.

Risk assessment is a concept that has developed to the point where it has the potential to address current limitations in cloud computing assessment methodologies. A risk assessment model for estimating the risk of cloud computing resources provides a solution to the risk problem, and would increase the chances of cloud computing adoption, as well as help in building trust in the cloud computing services. This thesis presents a new practical model of risk assessment to assess risk factors associated with cloud computing environment. In order to build a comprehensive risk assessment methodology, an extensive literature review was conducted to identify all risk factors that may affect cloud computing adoption. In this context 18 risk factors were identified. After the identification of risk factors, feature selection methods used to select the most effective features. The novelty of this thesis comes from the use of machine learning technique as a novel and efficient technique to assess risk in cloud computing

environment. To build the model; first data mining algorithms are applied, then the ensemble method is used to combine the outputs of the data mining algorithms.

The results of this research demonstrate the strengths of the use of data mining algorithms to assess risks, and it indicates that the methodology of using ensemble of machine learning algorithm represent a valuable alternative to existing methodologies.

ملخص

لقد برزت تقنية الحوسبة السحابية الالكترونية مؤخراً كصيغة جديدة لاستضافة وتقديم الخدمات للمستخدمين عبر الانترنت. علماً أن نظام الحوسبة السحابية هو عبارة عن مجموعة من الامكانيات المتاحة للاستخدام حسب الحاجة لتشغيل تطبيقات الحاسب الآلي بدلاً من تخصيصها لمجموعة من التطبيقات النمطية الجاهزة كما هو الحال فيما يخص العمل بنظام السيرفرات المرتبطة بخدمة العملاء. فقد تم التعريف بالحوسبة السحابية وتم تسويقها في ظل وجود عدد من المزايا المغرية لكثير من الشركات والمدراء في انحاء العالم المختلفة مثل انخفاض التكلفة والتخلص من عبء ضخم يتمثل في الإشراف على بنية تحتية معقدة من تأسيسات وتجهيزات تقنية المعلومات. علماً أن تقنيات الكيانات الافتراضية توفر امكانية تجميع الامكانيات والموارد في شكلها التجريدي بحيث تتم الاستفادة من الإدارات المختلفة في المؤسسة او الشركة. كما أن مراكز البيانات يتم تصميمها على اساس استخدام أجهزة ومعدات افتراضية وهي بمثابة الوحدات الذرية الجديدة للحوسبة.

كان هنالك اعتقاد تقليدي سائد وهو أن أي ربط بأنظمة أخرى أو جهات أخرى خارج المؤسسة أو الشركة من شأنه ان يفتح ثغرة لجهات لا تملك الصلاحية بأن تخترق المعلومات الخاصة بالجهة المعنية وتعبث بها. فيما يتعلق بالحوسبة السحابية يتم إبعاد البيانات بعيدا من جهاز الكمبيوتر المكتبي و الكمبيوتر المحمول ووضعها في مركز بيانات ضخم موزع على انحاء العالم المختلفة. وبالتالي، تبرز الحاجة إلى إجراءات فعالة لتقييم المخاطر من اجل احتواء كل انواع المخاطر.

ان تقييم المخاطر يعتبر فكرة نشأت وتطورت بالدرجة التي اصبحت لديها القدرة للتصدي لأوجه القصور الحالية في منهجيات تقييم الحوسبة السحابية. أن نموذج تقييم المخاطر المستخدم في تقدير مخاطر الحوسبة السحابية يوفر حلاً لمشكلة المخاطر ومن شأنه زيادة فرص الاقبال على استخدام الحوسبة السحابية بالإضافة إلى المساعدة في بناء الثقة في خدمات الحوسبة السحابية. هذا البحث يتضمن تقديم نموذج عملي جديد لتطبيق تقييم المخاطر من اجل تقييم عناصر المخاطر المرتبطة ببيئة الحوسبة السحابية. ومن اجل بناء منهجية شاملة لتقييم المخاطر، تمت دراسة البيانات المتعلقة بهذا الموضوع لحصر كل عوامل المخاطر التي قد تؤثر على استخدام الحوسبة السحابية. وتم حصر 18 عنصر من عناصر المخاطر. ثم بعد الانتهاء من تحديد عناصر المخاطر، تم في هذا البحث استخدام طرق اختيار العناصر لاختيار اكثر العناصر تأثيرا على بيئة الحوسبه السحابيه. ويتضمن هذا البحث اسلوب جديد لتقييم المخاطر وهو استخدام اساليب التنقيب في البيانات باعتبارها من أحدث وأفضل الاساليب في تقييم المخاطر في بيئة الحوسبة السحابية. من اجل بناء النموزج تم استخدام خوارزميات خاصة تستخدم في التنقيب في البيانات. وبعد ذلك تم استخدام طرق المجموعات المتحدة لدمج مخرجات هذه الخوارزميات الخاصة.

نتائج هذا البحث توضح قوة استخدام اساليب تنقيب البيانات في تقييم المخاطر و تبين أن منهجية استخدام طرق المجموعات المتحدة لدمج مخرجات خوارزميات تنقيب البياناتتمثل بديلاً ذا قيمة كبيرة للمنهجيات الموجودة حالياً .

Table of contents

List of Tables

List of Figures

Chapter one

Introduction

"If you don't actively attack the risks, they will actively attack you" [1]

Internet has been a driving force towards various emerging technologies. Cloud computing is one of the latest emerging internet-based technologies. Machines in the largest data centers can be dynamically provisioned, configured and reconfigured to deliver services in scalable manner [2, 3]. Cloud computing is lucrative to business enterprises. It provides the business with all the functionality of existing information technology services, and eliminates the to plan ahead for provisioning of resources. It allows enterprises to start with limited resources and increase resources only when there is a rise in service demand [4]. Cloud computing represents a fundamental change in the way Information Technology (IT) is created, evolved, deployed, scaled, updated, maintained and paid for [5]. The aim of cloud computing is to support the next generation data centers by architecting them as a network of virtual services. As such the users become able to access and deploy applications from any place around the world on demand at competitive costs depending on users QoS requirements [6, 7].

1.1 Problem statement

Cloud computing paradigm is targeted to provide a better utilization of resources using virtualization techniques and eliminates much of the client's work load. However, cloud consumers are afraid adopting cloud services technologies because of the lack of adequate confidence in cloud services in terms of the uncertainties associated with its level of quality. For instance, when using traditional technologies customer's software and data are stored in their computers, not like cloud computing technologies, which

moves the application software, and data to the large data centers, where the management of the data and services are not trustworthy. In terms of risk this poses many new challenges and risks that must be taken into account [8, 9]. In traditional architectures, the risk was enforced by an efficient security policy that addresses constraints on missions and flow among them, constraints on access by external systems and adversaries including programs and access to data by people. In cloud computing environment, this perception is totally obscured. The control in cloud computing environment is delegated to the infrastructure owner organization to implement sufficient policies that guarantee appropriate activities are being performed and ensure risk is reduced. This leads to a natural concern about data and asset safety, also it introduces additional number of new risks and threats that need to be assessed [10].

The importance of risk assessment in cloud computing environment is an outcome of the need to support different parties involved in decision making with respect to adopting cloud computing environment. Cloud service consumers are afraid adopting cloud services technologies because of the lack of adequate confidence in cloud services in terms of the uncertainties associated with its level of quality. An effective and efficient risk assessment of service provision and consumption, together with the corresponding mitigation mechanisms, may at least provide a technological assurance that will lead to high confidence of cloud service customers on one hand and a cost-effective and reliable productivity of cloud service providers resources on the other hand. Risk assessment in both terms of the process and techniques, offers an analytical and structured walk through of the organization's security state. Besides that, it outlines risk scenarios, identifies the consequences, should these occur, the frequency or likelihood of them occurring, the possible treatment options, and the associated costs [11].

Recently, there are a number of different types of risk assessment models, standards, and guidelines that are available; some of which are qualitative, others are more quantitative, while others are semi-quantitative (quantitative and qualitative). Each of these methods has been developed to meet a particular need and thus has different objectives, stages, structure, and level of application [12, 13]. Most of these researches works are for helping cloud consumers assessing risks before start using cloud computing and putting

their critical data in a security sensitive cloud. On the other hand it would help cloud providers assess and maintain risks in order to motivate consumers adopt cloud computing services. All these researches have laid a solid foundation for cloud computing. However, they hardly establish a complete risk assessment approach in consideration of a specific and complex characteristics of cloud computing environment, and they didn't used machine learning techniques to assess risk in cloud computing environment. There were neither complete quantitative nor qualitative risk assessment model for cloud computing. Therefore, cloud computing consumers need a new risk assessment to tackle the risks, and to check the effectiveness of the current security controls that protect an organization's assets. At present, there is a lack of risk assessment approaches for cloud consumers.

The need to rank and prioritize risks was generally mentioned in risk assessment literature; in order to identify areas for instant improvement and thus, concentrate the best efforts on minimizing the negative effect of risk events. With this aim in mind, we present a new risk assessment model, semi quantitative cloud risk assessment model, which has the main purpose of ranking cloud computing risks. Its main difference compared to other risk assessment models and frameworks in cloud computing, is that it evaluates the impact of risks using machine learning techniques. The intention was to address risks regarding the cloud computing environment and to provide a more structured, integrated, and inclusive model that provide necessary information required for the sustainable assessment of cloud computing environment risks.

1.2 Research Motivation

Though cloud computing is an innovative and promising paradigms that induce remarkable changes in the way in which hardware and software are designed and purchased, as well as how IT systems are managed. Cloud computing is a risky paradigm that is fraught with many risks. These risks can have a great impact on the operation of cloud providers, making it inconsistent with their respective business strategies represented by means of business objectives. On the other hand it prevents the consumers from adopting cloud services. Thus, both cloud service provider and cloud consumer

need proper risk assessment strategies to address and maintain these risks at an acceptable level. Appropriate risk assessment model can help cloud computing provider to maximize and win the trust of their consumers, and on the other hand help cloud computing consumers to be aware of the risks and vulnerabilities present in the current cloud computing.

1.3 Research Objectives

The focus of this research is to develop a reliable and effective risk assessment methodological tool for risk factors in cloud computing environment. In pursuit of this goal, the following specific research objectives were established:

1- To examine the body of knowledge about risk factors associated with cloud computing and identify them.
2- To simulate the dataset with regards to previously identified risk factors.
3- To develop a practical risk assessment model for cloud computing environment using machine learning techniques.

1.4 Thesis Structure

Chapter 2 begins by presenting a general overview about cloud computing, and discusses cloud emergence, definition and structure. Moreover, 18 risk factors associated with cloud computing are presented in this Chapter.

Chapter 3 defines terms and provides an overview of the concepts of risk, risk management, and risk assessment. In addition, it surveys existing literature review of risk assessment frameworks and methodologies in information systems in general specific to cloud computing.

Chapter 4 describes the methods and techniques used in thesis. The real implementation and experiments are discussed in Chapter 5 and the results of the methods are provided in Chapter 6. Finally, in Chapter 7 the conclusions are presented.

Chapter Two

Overview of Cloud Computing

This Chapter provides general over view about cloud computing environment. A background about cloud computing emergence is presented in Section 2.1. Since there is no standardized definition for cloud computing, Section 2.2 discusses the various definitions and Section 2.3 presents the cloud computing architecture.

2.1 Emergence of Cloud Computing

During 1990s, data center floor space, power, cooling, and operating expenses increased and lead to the adoption of grid computing and virtualization. Through grid computing users could plug in and use a metered utility service. The emergence of virtualization; by which the infrastructure be virtualized and shared across consumers, this motivated the service providers to change their business model to provide for remotely managed services and lower costs. Then, the wide distribution of the services lead to the need for integration and management of these services became important. All these technologies lead to the evolution of service-oriented architecture (SOA). Cloud computing developed out of this need to provide IT resources 'as-a-service' [14]. Figure 2.1 demonstrate the advancement of several technologies that lead to the emergence of cloud computing.

2.1.1 Cloud Computing Related Technologies

Autonomic computing was originally coined by IBM in 2001, it aims to build computing system capable of self-management. Unlike cloud computing, which aims to lower the resource cost rather than to reduce system complexity.

Virtualization is a technology that abstracts away the details of physical hardware and provides virtualized resources for high-level applications. The virtual machine forms the base of cloud computing, as it provides the capability of pooling computing resources from clusters of servers and dynamically assigning or reassigning virtual resources to applications on-demand.

Grid computing is a distributed computing paradigm that coordinates networked resources to achieve a common computational objective. Cloud computing is similar to grid computing in that it also employs distributed resources to achieve application-level objectives. However, cloud computing has one step further by leveraging virtualization technologies at multiple levels to realize resource sharing and dynamic resource provisioning.

Utility computing represents the model of providing resources on-demand and charging customers based on usage rather than a flat rate. cloud computing is the realization of utility computing. With on-demand resource provisioning and utility-based pricing, service providers can truly maximize resource utilization and minimize their operating costs [4].

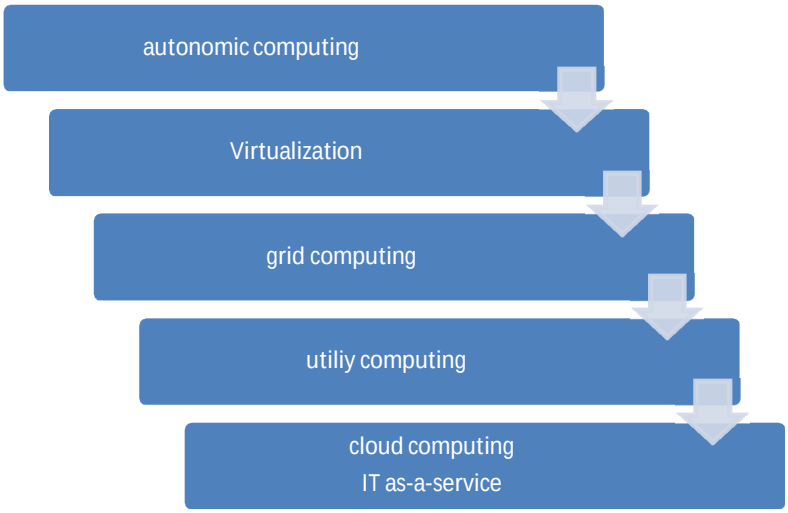

Fig 2.1. Various technologies lead to cloud computing emergence

2.2 Cloud Computing Definition

The main idea behind cloud computing is not a new one. Cloud computing is a conglomerate of several different computing technologies and concepts like grid computing, autonomic computing, service oriented architecture, virtualization, peer-to-peer computing (fig. 1). Cloud computing has inherited many of these technologies benefits and drawbacks [15]. John McCarthy in 1960, envisioned that computing services will be provided to the general public like a utility [16]. In 1969, [17] said: “as of now, computer networks are still in their infancy, but as they grow up and become sophisticated, we will probably see the spread of ‘computer utilities’ which, like present electric and telephone utilities, will services individual homes and office across the country.” The term “cloud” has also been used in various contexts to represent many different ideas, this lack of a standard definition of cloud computing generates many issues such as: market hypes, and a great amount of skepticism and confusion [4]. Recently, there has been work in standardizing the definition of cloud computing, and many practitioners in the commercial and academic fields have attempted to define exactly “what cloud computing” is and what associated characteristics it represents.

The work in [18] compared more than 20 various definitions come from different sources to confirm a standard definition. [5, 7, 19] has defined cloud computing as “cloud is a parallel and distributed computing system consisting of a collection of interconnected and virtualized computers that are dynamically provisioned and presented as one or more unified computing resources based on service-level agreement (SLA) established through negotiation between the service provider and consumers”. The National Institute of Standards and Technology (NIST) [20] defined cloud computing as: “Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or

service provider interaction". The main reason for the existence of many definitions of cloud computing is that cloud computing is not new technology, but rather a new operations model that brings a set of existing technology together to meet the technological and economic requirements of today's demand for information technology [4]. This cloud model is composed of five essential characteristics, three service models, and four deployment models.

2.3 Core Technologies

To better understand the risks associated with cloud computing, it is important to discuss the core concepts and technologies in cloud computing. Cloud computing realize the dream of utility computing, which means that all computing services can be provided in a similar manner to the other utilities such as electricity. To provide hardware services as measured service-oriented can be easily understood, but this can also be extended to software services because they are designed and built in the form of autonomous interoperable services. The ways to access cloud computing were expand due to the large variety of devices that can connect the Internet. Data centers, server farms, and high-speed broadband networks are also critical components.

Virtualization is the most effective technology in cloud computing. Virtualization means, to create a virtual versions of computers or operating systems. By the use of virtualization, all physical traits are hidden from the user and instead another abstract computing platform is presented, a key concept in cloud computing [21].

2.4 Cloud Computing Architecture

Many organizations and researches have defined the architecture for cloud computing [4, 20-22]. Cloud architecture is the design of software applications that uses internet-accessible on-demand service. In this Section, we describe cloud computing architecture based on NIST [20] definition.

2.4.1 Cloud Computing Layers

Cloud computing consist of four layers, as shows in Fig. 2.2. We describe each of them in detail:

Hardware layer*:* It includes router, switches, physical servers, power and cooling system. Hardware layer is responsible for managing the physical resources of the cloud and is implemented in data centers.

Infrastructure layer (virtualization): Virtualization technologies made many key features such as dynamic resource assignment available, thus, the infrastructure layer become an essential component of cloud computing. Infrastructure layer partition the physical resources using virtualization technologies such as KVM [23], Xen [24], VMware [25], to create a pool of storage and computing resources.

Platform layer: Consists of operating systems and application frameworks is built on top of the infrastructure layer. The platform layers main objective is to decrease the load of deploying applications directly into VM containers.

Application layer: Consists of the actual cloud computing applications, which can leverage the automatic-scaling feature to achieve better availability, performance, and lower operating cost.

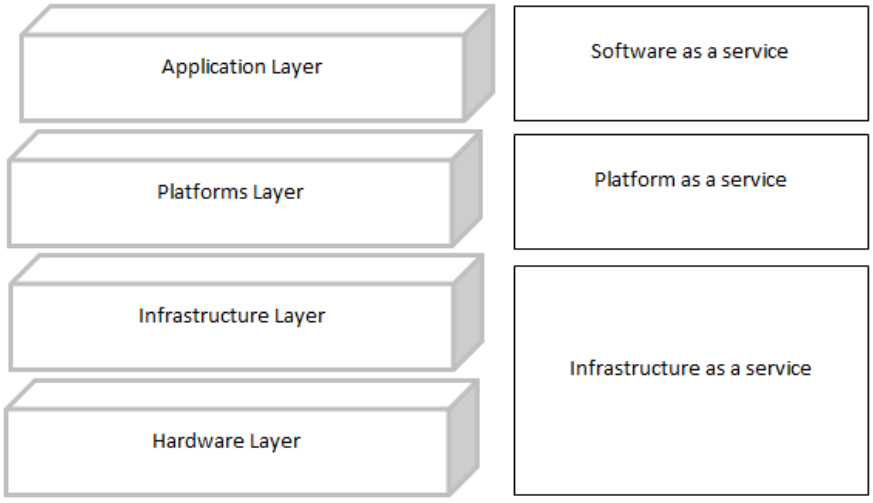

Figure 2.2 cloud computing layers and service models

2.4.2 Cloud Computing Service Models

Cloud services are delivered and consumed in real-time over the Internet. There are three categories of cloud services: infrastructure, platform, application.

1. **Infrastructure as a service (IaaS):** This model provides computer resources, storage, and network as an Internet-based service and is based on virtualization technology. The key benefits of IaaS is the usage-based payment strategy which allow customers to pay as they grow, it keeps the customers use the latest technology always, and achieve a much faster service delivery and time to market. The cloud owner who offers IaaS is called an IaaS provider. The most familiar IaaS provider is Amazon EC2.
2. **Platform as a Service (PaaS):** The cloud provider delivers with a platform including tools, and all the systems and environments comprising the end-to-end lifecycle of developing, testing, deploying, hosting, and manage their own applications as a service, and without installing any of these platforms or support tools on their local machines. The platform-as-a service strategy can minimize development time, offer hundreds of readily available tools and services, and quickly scale. The PaaS model may be hosted in to of IaaS model or on top of cloud infrastructure immediately. Microsoft Azure, and google Apps are key examples of PaaS.
3. **Software as a service (SaaS):** Cloud provider deliver applications hosted in the cloud infrastructure as internet-based services for end users without requiring installing the applications on the customers computers. SaaS is a multi-tenant platform. It uses common resources and a single instance of both the project code of an application as well as underlying database to support multiple customers simultaneously. This model can be hosted on top of PaaS, IaaS, or hosted on cloud infrastructure directly. Examples of key providers of SaaS are sales force, Microsoft, and IBM.

The development of standard security model for each service delivery model is difficult, because each service model has different possible implementations. Furthermore, the existence of all service delivery models in one cloud platform leading to further complication of the security management process. Figure 2.3 illustrates a simple cloud computing model.

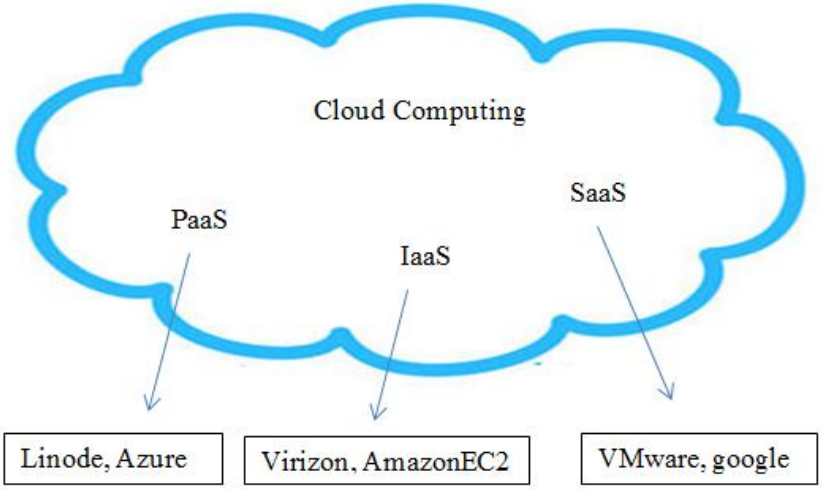

Figure 2.3 Illustration of a cloud computing model with its service models

2.4.3 Cloud Computing Deployment Models

Depending upon cloud computing customers' requirements and on services characteristics and purpose; cloud services can be deployed in four ways each with its own benefits and drawbacks. The deployment models include public(external),private (internal), community, hybrid clouds, and virtual private clouds. These models are discussed below:

Public cloud: Cloud providers provide their resources such as storage and applications as services to the general public, via a web application or web services over the internet. The resources are therefore located at an off-site location that is controlled and managed by the service provider or a third party. These are typically low-cost or pay-on-demand and highly scalable service. The key benefit of public cloud is the shifting of risks to infrastructure providers. However, public clouds lack control over date, network, and security settings which obstruct their effectiveness in many business projects.

Private clouds: Cloud service in this model is made available to specific customer and may built and managed either by organization itself or third party and may exist at an on-site or off-site location. The key benefits of private clouds that it offers the highest degree of control to the provider and user over cloud infrastructure, improving performance, reliability, compliance, transparency and security. In the other hand, private clouds require capital expenditure, operational expenditure and highly skilled IT team, and it often being similar to traditional proprietary server farms and do not provide benefits.

Community clouds: They are controlled and shared by several organizations and support a specific community that has shared interests, such as mission, policy, security requirements and compliance consideration. It may be managed by the organizations or a third party and may exist at on-site or off-site locations, and the members of the community share access to the data and applications in the community cloud. Community clouds users therefore seek to exploit economies of scale while minimizing the costs associated with private clouds and the risks associated with public clouds.

Hybrid clouds: A combination of two or more clouds (private, community, or public) that remain unique entities but are bound together by standardized or propriety. Hybrid model tries to address the limitations of public and private cloud by making a combination of them. Applications with less stringent security, legal, compliance and service level requirements can be outsourced to the public cloud, while keeping business-critical services and data in a secured and controlled private cloud. The advantage of hybrid clouds over private and public cloud that it offer more flexibility. certainly, they provide more strict control and security when compared to public clouds, while still facilitating on-demand service expansion and contraction.

Virtual private clouds [14]: This deployment model is described by fewer resources, is one in which service providers utilize public cloud resources and infrastructure to create a private or semi-private virtual cloud (interconnecting to internal resources), usually via virtual private network (VPN) connectivity.

2.4.4 Cloud Computing Essential Characteristics

Cloud computing provides several features that are different from traditional service computing. The national Institute of Standard and Technology (NIST) [20] addressed five characteristics as depicted below:

- On-demand self-service: cloud computing employs a pay-per-use pricing model. Resources can be allocated and deallocated on-demand and consumer can unilaterally provision computing capabilities as need automatically without requiring human interaction with each service provider.
- Broad network access: cloud service are available over the network (usually internet) and accessed through standard mechanisms such as mobile phones, laptops, and work stations.
- Resource pooling: the cloud providers use multi-tenant model to provide multiple consumers by their pooled computing resources. Physical and virtual resources dynamically assigned and reassigned depending on consumer demand.
- Rapid elasticity: resources can be scaled up and down rapidly and elasticity.
- Measured service: the resource usage controlled and optimized by leveraging a metering capability.

2.5 Cloud Computing Risk Factors

The evolution of cloud computing paradigm introduced new risks, specific issues imposed by law or regulations, as well as operational risks inherent to the use of cloud systems, either local or external assets. These risks can have a great impact on the operation of cloud providers, making it inconsistent with their respective business strategies, represented by means of business objectives and constraints. Risk in cloud computing systems must be considered at service, data, and infrastructure layers. The type of cloud computing environment affect the level of risk. For instance, being a participant of a federation of clouds involves more risks than only managing a private cloud. Cloud computing literature review is navigated and a lot of risk factors is founded. Some issues is observed: many researches define same risk factor but they use under different names; other researches define risk factor but it can be included in another one.

Thus, with regards to these issues 18 risk factors is identified. This section give a detailed definition about each risk factor that may affect adoption to the cloud computing environment.

2.5.1 Authentication and Access Control (A&AC)

Cloud computing customer's private and sensitive data must be secure and only authenticated users can access it. When using cloud, the data is processed and stored outside the premise of an enterprise, which brings a level of risk because outsourced services bypass the "physical, logical, and personnel controls", any outside or unwanted access is denied. The level of access control could enable attackers to collect confidential data or to gain complete control over the cloud services. Malicious insider attack which can be performed by malicious employees at the provider's or user's site. A malicious insider attacker can easily obtain passwords, cryptographic keys and files. Account or service hijacking can take place due to unauthorized access gained by attackers, it can happen if the attacker gain access to user's credentials then he or she can eavesdrop on its activities, transaction, manipulate data, and redirect the customer to illegitimate sites. Abuse and nefarious use of cloud attack can happen because some cloud providers offer free limited trial periods which enables hackers to access the cloud immorally, some malicious hacker use cloud server to launch DDoS attack, propagate malware or share thieved software. All this ways of authentication access can cause damage and financial productivity losses, their impact appears on: the confidentiality, integrity, and availability of all data. The cloud provider should have their own identity management to identify individual and controlling the access to the resources, authentication and authorization through the use of roles and password protecting is a common way to maintain access control to a cloud computing systems [8, 26-28].

2.5.2 Data Loss (DL)

Data loss means that the valuable data disappear into the ether without a trace. Data may loss by several ways such as: some malicious hackers may delete or alter records without having backup, Some customers may encrypt their data to prevent theft, but this can be backfire if they lose the encryption key, which can be very painful, unauthorized access,

operational failures, unreliable data storage, unlinking record, and sometimes due to the careless of cloud service provider or a disaster, such as earthquake, flood, fire. Cloud customers need to make sure that this will never happen to their sensitive data. This risk is take place because of the amount of data access operations and the kind of data stored on clouds. Data loss is one of the top concerns for business, because this may lead to lose their reputation, and cause a loss that may significantly impact customer morale. There is a need for data leakage security to implemented so, the important data will not go into the wrong hands, good access control must be taken into practice, data must be stored securely and integrity, periodically, monitoring must be taken into practice [2, 26, 29-32].

2.5.3 Insecure Application Programming Interfaces (IAP)

Application Programming Interfaces (APIs) are software that provided by cloud service provider for customers to use to manage and interact with their services. APIs need to be secured because they are important and necessary part to security and availability of whole cloud services, and they play an integral part during provisioning, management, orchestration and controlling cloud computing environment processes. Building interfaces, injecting services will increase risks, there for some organization may in force to relinquish their credentials to third party in order to enable their agency. Different security issues may expose if the interfaces is comparatively weak, security control mechanisms may not be able to fend API hacks, which may lead to unauthorized access to even privileged user functions, and if the cloud providers provide some kind of software interfaces to a customer to manage and interact with their services, the week or too much user friendly interfaces may generate different kinds of security issues, and sometimes. The risk increased in customer management interfaces of public cloud because they are Internet accessible and mediate access to larger sets of resources especially when combined with remote access and web browser vulnerabilities. The security and availability of cloud services is associated with these APIs so they should include features of encryption, access control, authentication, and activity monitoring [26, 29, 31, 32].

2.5.4 Data Transfer (DT)

Sensitive data is obtained from customers, processed and stored at cloud provider end. Security of the data leaving a data-center to another data-center is a major issue as it may breached quite a number of times in the recent time. All data flow over network needs to be secured in order to prevent seepage of customer's sensitive information. The application provided by cloud provider to their customers is has to be used and managed over the web. The risk come from the security holes in the web application [8, 33].

2.5.5 Insufficient Due Diligence (IDD)

Cloud computing come with the promise of cost reductions, operational efficiencies and improved security. While these can be realistic goals for organizations, too many enterprises jump into the cloud without understanding he full scope of it. Before start using cloud services, the organization need to fully understand the cloud environment and its associated risk. An organization must be sure that they have appropriate resource and they have a team that are familiar with cloud technology to prevent the issues may arise from jumping to cloud computing such as operational and architectural issues. Beside that an organization need to be sure about adequate performance, because if moving to cloud will save money. However if the performance level is unacceptable then no need to this saving, also when making application testing an organization should validate that any operations with cloud storage service work correctly, and the tester should be aware that cloud services often perform much slower than local services. Organizations must taking on unknown levels of risk in ways they may not even comprehend, but that are a far departure from their current risks [32, 34].

2.5.6 Shared Environment (ShE)

Multi-tenancy is key factor of cloud computing service. To achieve scalability cloud provider provide shared infrastructure, platform, and application to deliver their services, this shared nature enable multiple users to share same computer resources, which may lead to leaking data to other tenants. The key is that a single vulnerability or misconfiguration can lead to a compromise across an entire provider's cloud. Moreover, the concept of the isolation of the individual cloud users does not have a sure

implementation in the cloud environment. If one tenant carried malicious activities the reputation of other tenants may be affected. The impact can be appear as a problems for the organization's reputation in addition to service delivery, and data loss. Also one flaw could allow an attacker to see all other data. If the foundation of computing resources not offers strong isolation for a multitenant, the risk arises in all delivery models. Related to shared access is data confidentiality or privacy risk arise. To control shared environment risks, cloud provider should monitoring the environment for unauthorized activity, and must conduct vulnerability scanning and configuration audits [2, 29, 32, 34, 35].

2.5.7 Regulatory Compliance (RC)

Traditional service providers are subjected to external audits and security certification, and they give their customers some information about the security controls that have been evaluated. European Economic area (EEA) has enacted data protection laws requiring that the obligation to provide adequate data security should be passed down to subcontractors, many others countries have been passed similar laws. They establish that the provider is responsible for ensuring the protection, security and integrity of the data regardless of location, also the provider remain liable for any loss, damage or misuse of the data. Any provider is unable or unwilling to undergo such audit should only be considered for most trivial functions [28, 35, 36].

2.5.8 Data Breaches (DB)

Virtual machine (VM) could use side-channel timing information to extract private cryptographic keys in use by other VMs on the same server. Cloud environment present a high value target to attackers, and therefore, the data from different users hosted in cloud environment. Breaching in to cloud environment will potentially attack all users data. Those attackers can exploit a single flaw in one client's application to get to all other client's data as well, if the cloud service databases are not designed properly. Besides that, there is a risk from insiders, although they don't have a direct access to databases, the insider breaches risk is still high and can be a massive impact on the security [8, 32].

2.5.9 Business Continuity and Service Availability (BC&SA)

Cloud providers business continuity and service availability is essential issue especially for critical business process, organizations worry about whether utility computing services will have adequate availability, and this make some warry about cloud computing. The continuity and availability of service refers to the factors that may negatively affected the continuity of cloud computing. The nature of business environment, competitive pressure, and the changes happening in it leads to some events that may affect the cloud service provider, such as merger, go broke, bankruptcy, or it acquisition by another company. These things lead to loss or deterioration of service delivery performance, and quality of service. Another important thing to the cloud computing provider is that their customers must be provided with service around the clock, but outages do occur and can be unexpected and costly to customers. Cloud service availability can be affected by many factors such as natural disaster, which can cause cloud services to become unavailable or lead to loss of Internet connectivity. Another factor that may affect availability is the priority of users on the cloud, how it determined, should the overcapacity threshold is reached. The dependence of organization on 24/7 availability on some services increases the problems with Denial of Service (DoS), failure can cost service providers and customers. Cloud services exploited by cyber criminals to make distributed denial-of-service (DDoS) attacks, resulting in flood a web server with repeated message causing hanging up the system and denying access for legitimate users [3, 8, 37, 38].

2.5.10 Data location and Investigative Support (DL&IS)

Most cloud service providers have many data centers around the globe. When the customer start using the cloud platform, they are not aware about the place of the datacenter in which their data stored beside that they don’t have any control over the physical access mechanisms to that data. When regards to privacy regulation in different jurisdiction, in different countries where the government restrict the access to data in their borders, or if the data stored in high-risk countries, all these things make data location big concern issue. The investigation of an illegal activity may be impossible in cloud

computing environment, because multiple customer's data can be a located in different data centers that are spread around the globe, which makes the investigation difficult, time consuming, and expensive. The enterprise has to factor in the inability or unwillingness of the provider to support the processing of business records or anticipates the need for investigation [8, 28, 35]**.**

2.5.11 Data Segregation (DS)

Multi tenancy and shared resource are major characteristics of cloud computing where multiple users can share same computing capacity, storage, and network. All cloud providers use secure sockets layer to protect data in transit. The risk arise here come from the failure of the mechanisms to separate data in storage, and memory, from multiple tenants in the shared infrastructure. To observe system and end user security behaviors, the existence or absence of technical issue such as encrypted communication and virtualization security, and fundamental architectural concerns such as a dependence on the Internet and missing choke points can be used. In this environment the intrusion of data of one user becomes possible, therefore, the probability of this scenario depend on cloud model the likely in private models is lower than public models [8, 28, 35, 36, 39].

2.5.12 Recovery (R)

ENISA (2009) finds that 52.8% of SME (Small and Medium Enterprise) vote disaster recovery capabilities as a reason for start using cloud computing. Cloud users do not know where their data is hosted. Some events such as man-made, or natural disaster may happen; in such events customers must require information on what happens to their data in case of disaster and how long the recovery process take [8, 28, 35].

2.4.13 Virtualization Vulnerabilities (VV)

Virtualization is one of the fundamental features of the cloud service, due to it the cloud providers are residing the user's applications on virtual machine (VMs) in a shared infrastructure. The VMs is a virtualized based on the physical hardware of cloud provider. The cloud providers isolate the VMs from each other, due to security concerns. Hypervisor is the main source of managing a virtualized cloud platform, and it resides

between VMs and hardware. Cloud providers use it to provide virtual memory as well as CPU policies to VMs. Hypervisor introduces major risks as every cloud provider uses it, hackers targeted it to access the VMs and physical hardware, attack on hypervisor can damage the VMs and hardware. Besides, all virtualization software has vulnerability, which can be exploited by malicious, local users to bypass certain security restrictions or gain privileges. Strong isolation should be applied to ensure that if any VM is malicious, it will not affect other VMs under same cloud provider [8, 26, 37].

2.5.14 Third Part Management (TPM)

There are many issues in cloud computing related to third party because the client organizations are not directly managed by the cloud service provider. Some old concerns in information security appear with outsourcing such as integrity control and sustainability of supplier and all risks that client may take if it rely on a third party. In some situation, the level of security of the cloud provider may depend on the level of security of each one of the links and the level of dependency of the cloud provider on the third party, which may lead to supply chain failure risk, which can take place if cloud provider can outsource certain specialized tasks of its 'production' chain to third parties. Some issues can happen as a result of the lack of coordination of responsibilities between all the parties such as loss of data confidentiality, integrity and availability, unavailability of service, violation of SLA, economic and reputational losses due to failure to meet customer demand, cascading service failure, etc. Lack of transparency in the contract can be a problem for the whole system. Its impact can appear in decreasing the level of trust in the provider [39].

2.5.15 Interoperability and Portability (I&P)

In some cases the organization may need to change the cloud provider, and there have been cases when companies can't move their data and applications if they want to change the provider. Also, in some cases, the organization want to use different platforms for their applications. The difficulty of changing cloud provider (portability) or extracting data and programs (interoperability) from cloud provider is preventing some organization from adopting cloud computing. Thus, the organizations need to maintain a balance to

handle the interoperability and portability. Interoperability means the ability of systems to communicate, in other words is the ability of the code to run with more than one cloud provider simultaneously. Portability is the ability to run systems written for one environment in another environment. Interoperability and portability become crucial because if the organization locks to a specific cloud provider, then the organization will be at the mercy of the service level and pricing policies of that provider and it hasn't the freedom to work with multiple cloud provider. One solution would to standardize the APIs thus the SaaS developer could deploy services and data across multiple cloud providers so that if one company fails this would not affect all copies of customer data with it [33, 40].

2.5.16 Resource Exhaustion (RE)

Cloud provider allocates resource according to statistical projections. Inaccurate modeling of resources usage can lead to many issues such as: service unavailability, access control compromised, economic and reputational losses, and infrastructure oversize [36].

2.5.17 Service Level Agreement (SLA)

This term exists in different applications including the cloud. SLA is an agreement between a service provider and a service customer, it specifies the responsibilities of the service provider and the customer, besides, the information about the service delivered by the cloud provider, the QoS provided, in addition, the penalties if the contract terms are broken by the cloud provider [41, 42]. In other words the (SLA) represents the foundation for the costumer to trust in the provider. The organization needs to ensure that the terms of (SLA) are being met. Risk may appear with service level application such as the data owner as some cloud provider include explicitly some terms state that the data stored is the provider's not the customer's. If the cloud vender is owing the data it gives them more legal protection in case if something goes wrong, beside that they can get additional revenue opportunities for themselves by searching and mining customer data [31]. In few cases where cloud vendor went out of business, their customer private data sold as part of the asset to the next buyer. Also (SLA) terms should include Licensing conditions, there

is the possibility for creating original work in the cloud, but if not protected by the appropriate contractual clauses, this original work may be at risk. One of the (SLA) terms must be for responsibilities of cloud provider for enabling governance [43].

2.5.18 Data Integrity (DI)

One of the most critical elements in all systems is data integrity. It is easy to achieve in a standalone system with a single database and it can maintain via database constraints and transactions. Achieving data integrity is much complex in distributed systems where there are multiple databases and multiple applications. Cloud computing magnified the problem of data integrity, as there is mix of on premise and SaaS applications exposed as service. SaaS applications are multi-tenant applications and they hosted by a third party. The biggest challenge, which endanger the data integrity is transaction management, at the protocol level, does not support transactions or guaranteed delivery. If data integrity is not guaranteed and there is lack in integrity controls, this may result in deep problems [8].

2.6 Summary

This chapter considered cloud computing and discussed the emergence, definition, and the core technologies of cloud computing. We also presented cloud computing architecture (layers, service models, deploy models) and the various risk factors, which affect cloud computing adoption.

Chapter Three

Risk assessment and Literature Review

During the last decade there has been a major surge of interest in improving our ability to deal with risk, and especially with its negative impact at the organization level. This has led to the development of tools, techniques, processes and methodologies which are typically classified under the label of "risk management" [44]. This chapter examines the definition of risk in Section 2.1. A general overview about risk management and risk assessment is provided in Sections 2.2, and 2.3 respectively. Some basic background about risk assessment approaches is presented in Section 2.4.

3.1 Risk Definition

Risk is a part of any activity and can never be eliminated, nor can all risks ever be known. Risk in itself is not bad; risk is essential to progress, and failure is often a key part of learning. But we need to learn how to balance the possible negative consequences of risk against the potential benefits of its associated chance [45]. There are a variety of alternatives but accepted definitions exist for the term risk, these definitions depend broadly on the disciplines within which the concept is applied. [46] proposed a generic definition of risk "the probability that a particular adverse event occurs during a stated period of time or results from a particular challenge". This definition provides a useful guide to the general meaning of risk, but it does not meet the more quantitative requirement of the engineering profession, or more qualitative requirement of social science disciplines. For this reason there is a general acceptance that definitions of risk are in general case specific [46]. ISO 31000:2009 [47] together with ISO/IEC Guide 73 [48], defines risk as the "effect of uncertainty on objectives". It also states that risk is consequence of an organization setting and pursuing objectives against an uncertain environment. Risk is not defined or classified by the size of the risk, by the balance of

expected and unexpected consequences, which is known as “value at risk” in economic terms. “value at risk” is statistical measure that defines the consequence of a loss by the chance of occurrence or confidence level [49]. Usually, risk is defined as the combination of severity and probability of an event. In other words, how often can it happen and how bad is it when it does happen?. Risk can be evaluated qualitatively or quantitatively.

$$Risk = frequency\ of\ the\ event * consequence \quad (3.1)$$

The terms can be used in the qualitative descriptions of risk such as ‘low’, ‘medium’, or ‘high’. The terms can be used in quantitative descriptions of risk is numeric values such as ‘1’, ‘2’ [50].

Risk can be classified into three broad categories [50]:

1. Negligible risk: broadly accepted risks as they considered they go about everyday lives.
2. Tolerable risk: we would rather not have the risk but it is tolerable in view of benefits obtained by accepting it. The cost is in convenience is balanced against the scale of risk and a compromise is accepted.
3. Unacceptable risk: the risk level is so high that it couldn’t be prepared to tolerate it. The losses far out weight any possible benefits in the situation.

The discussion of risk is come from its relationship with the idea of reward. If the risk is not associated with well understood or widely-accepted cost, the organization faced a challenge. Because of this fail in managing risk there is a need to develop risk management programs in order to identify, mitigate, and manage risks to achieve acceptable reward [49].

3.2 Risk Management

Risk management is “the process of understanding, costing, and efficiently managing unexpected levels of variability in financial outcomes for business” [49]. Risk management plays an important role in a wide range of fields, including statistics, economics, systems analysis, operation research, and biology [51]. The aim of risk

management is to help organizations establishing priorities and focusing security resources in order to reduce risk exposure [52]. Risk management process aim is three fold: it must identify the source of uncertainty, assess the frequency of events occurrence and consequences of those events, and respond to the risk in an appropriate and effective manner [53]. The risks associated with information, information system, and technology are included in any definition of commercial or public-sector risk. System risks are potential system losses, breaches, or failures which may mean "modification, destruction, theft, or lack of availability of computer assets such as hardware, software, data, and services" [54]. Further, risk management is the process that allows IT managers to balance the operational and economic costs of protective measures and achieve gains in mission capability by protecting the IT system and data that support their organization mission [55]. The most central concepts in risk management are the following [51, 52]:

(a) Assets: is something to which a party assigns value and hence for which the party requires protection; they are not only hardware, networks or software, but also all those supporting the underneath infrastructure such as staff or facilities.
(b) Threats: is a potential cause of an unwanted incident, with unwanted results for an organization's objectives materialized on harm or loss of assets.
(c) Vulnerabilities: is a weakness, flaw on procedures, design, implementation or internal security controls in IS, that may be exploited purposely or accidentally by, a threat to cause harm to or reduce the value of an asset.
(d) Risk: it is the potential that a given threat will exploit a vulnerability of an asset and thereby cause harm to the organization. It measured by the likelihood of an unwanted incident and its consequence for a specific asset.

Risk management is an iterative process and the identified risks are monitored throughout the lifecycle, it not concerned about eliminating risk but about identifying, assessing, and managing risk. Its main goal is to obtain benefits and sustainable values for the business in each of its activities and across all of them. For this reason, it should be a fundamental part of any organization's strategic management [56]. Figure 3.1 shows the steps of risk management process.

Fig 3.1 Risk Management Steps

3.3 Risk assessment

Risk Assessment (RA) is a set of techniques applied in order to investigate the probability of an event, and there by assess the effects/consequences of such. It is one element of a broader set of risk management activities. Although all elements of risk management cycle are important, risk assessment provides the foundation for other elements in the cycle, it considered a core sub process of any risk management strategy: if the risk assessment method is not conducted appropriately, the risk management will then fail to achieve its objectives. In particular, risk assessment provide a basis for establishing appropriate policies and selecting cost-effective techniques to implement these policies [57]. Risk assessment can be defined as “the process that tries to identify, analyze, and evaluate through abroad range of involved variables, potential events with a measurable impact on an organization’s objectives [47]. Risk assessment, regardless is related to any type of risk it provides risk-level estimations (RLEs) as output, and it is considered as a means of providing decision makers with information needed to recognized factors that can negatively influence operations and outcomes and make informed judgements concerning the extent of actions needed to reduce risk [57]. The complexity of risk assessment grows along with the environment and information system complexity. For

RA to be further useful, it must be precise and allows contrast and comparison against previous assessments, or against assessments done in similar environments [52].

Risk assessment meaning largely depends the context and discipline within which it is applied, but all risk assessment process can be subdivided into: identification, analysis, and evaluation of risks [56, 58]. Risk assessment steps are shown in figure 3.2

3.3.1 Risk identification

This stage establishes and defines an organization's potential events and their causes and potential consequences, understanding the source of risks, and areas of impact. The risks were described in structure format in this step. The goal is to create a comprehensive list of risks, including risks that may be associated with missed opportunities and risks out of the direct control of the organization. It considered as the most difficult aspect of managing risks, because more events will happen in the future than can be predicted today.

3.3.2 Risk analysis

This stage aimed at figuring out the everything possible about risks, including events likelihood of occurrence and the magnitude of their consequences, based on previously identification risks. Existing controls and their effectiveness and efficiency are also taken into account. There are three types of risk analysis: quantitative, qualitative, and combination of two.

a. Quantitative risk analysis

Attempts to estimate the risk in the form of the frequency of the events and the magnitude of the losses or consequences. Quantitative risk analysis is the most suitable method when sufficient field data, test data or other evidences exist so as to estimate the probability of events and magnitude of losses; however, it is complicated, time consuming and expensive to conduct.

b. Qualitative risk analysis

It is main features that it is simple and quick to perform; thus, it is the most widely applied method. Linguistic scale such as low, medium, and high is using to estimate risk; then, a matrix is formed, which characterizes the risk in the form of the likelihood of events versus the potential magnitude of losses in qualitative risk analysis does not rely on actual data and probability treatment of such data; accordingly, it is more simpler to use and understand than quantitative risk analysis, although it is extremely subjective.

c. Mixed risk analysis

It adopts a combination of quantitative and qualitative analyses. It can be happen in two ways: either the frequency of an event is measured qualitatively, but the consequences are measured quantitatively or vice versa; or both the frequency of an event and the consequences are measured using quantitative methods, but the policy setting and decision making are reliant on qualitative methods.

3.3.3 Risk evaluation

This stage compares the estimated risk levels against a risk acceptance criterion, which is a threshold established by business executives. The purpose of risk evaluation step is to review the analyses, criterion and tolerance of risks in order to prioritize and choose appropriate risk treatment methods. An organization's legal and regulatory environment and its internal and external context will also be considered at this stage. This step results in the determination of whether each risk should be treated or not, and how to treat it.

Fig 3.2 Risk Assessment Steps

The selection of an assessment technique is not a simple task. Authors in [59, 60] listed some issues should be put in considerable when selecting the most suitable assessment technique:

- The availability of resources for analysis.
- The size and complexity of the process analyzed.
- The phase in which the risk assessment will be considered in the process lifecycle.
- The availability of information.

The authors also emphasize the importance of the data considered in the risk assessment. The data considered should be accurate, adequate, relevant, coherent, unbiased and valid.

3.4 Literature Review

In the current information age, the issue of information security has become a vital entity. The evolution of cloud computing paradigm introduced new risks, specific issues imposed by law or regulations, as well as operational risks inherent to the use of cloud systems, either local or external assets. These risks can have a great impact on the operation of cloud providers, making it inconsistent with their respective business

strategies, represented by means of business objectives and constraints. Information security risk assessment enables the Government, public and private organization to identify their security risks, and develop effective control strategies. Nowadays, there are different types of risk assessment standards, methods, and guidelines are available. The different types of risk assessment have different objectives, structure, steps, and level of application; and each of them has been developed to meet a particular need. Several researches have been done in the area of risk assessment of information security in general and risk assessment in cloud computing environments. Some of these researches tend to provide new risk assessment models or frameworks to overcome potential risks associated with cloud computing and prevent adoption to cloud computing. The risk assessment strategies driven by business aspects must be integrated into a cloud organization's decision-making processes in order to be effective. Risks in cloud systems must be considered at service, data, and infrastructure layers. Besides that, all cloud entity not just the providers, should be the subject of risk assessment approach [56]. This research focuses on assessing risk in cloud computing environments. The following parts presents a detailed literature review about different researches, frameworks, approaches, and methodologies that has been done in the area of assessing risk in information security, and in the area of cloud computing environment as specific.

3.4.1 Risk Assessment Standards and Methodologies

There are many risk assessment standards, these standards are published by professional organizations in terms of information security as general, some are publicly available (e.g. OCTAVE), while others are restricted to members of organizations that are collaborating to create and update them (e.g. SPRINT). The following are the descriptions of each of these standards:

OCTAVE is a self-directed risk-based strategic assessment and planning technique for security. It requires an organization to manage the evaluation process and make information-protection decisions. OCTAVE dedicated at organizational risk and strategic, practice-related issues. This make it differ from the typical technology-focused assessment, which is focused at technological risk and tactical issues. OCTAVE structure

is designed in three phases figure 2.3, theses phases are: phase 1: Build Asset-Based Threat Profiles- in this organizational evaluation the analysis team determines what is important information-related-asset to the organization, then the team selects the most important assets and describes security requirements for each asset. Finally, they asset profile by identifies threats for that asset. Phase 2: Identify Infrastructure Vulnerabilities- this is an evaluation of the information infrastructure. First the analysis team examines network access paths, identify classes of information technology components related to each asset, at last, the team determines the extent to which each element of component is resistant to network attacks. Phase 3: Develop Security Strategy and plans- during this evaluation, the team identifies risks association with organization's assets and decides the action to do. The team used gathered information to create a protection strategy mitigation plans to address the risks to the critical assets [61].

A voluntary private-sector initiative called Internal Control-Integrated Framework was published in 1992, by the committee of Sponsoring Organizations of the Treadway Commission (COSO) [62]. In 2004 COSO published an Enterprise Risk Management (ERM) standard [63] which is sponsoring by five nonprofit organizations, American Accounting Association (AAA), American Institute of Certified Public Accountants (AICPA), Financial Executives International (FEI), Institute of Internal Auditors (IIA), and Institute of Management Accountants (IMA). COSO ERM is a multilayer project targeted to improve organizational performance and governance through effective internal control, enterprise risk management, and fraud deterrence. Figure 3.3 illustrates OCTAVE phases.

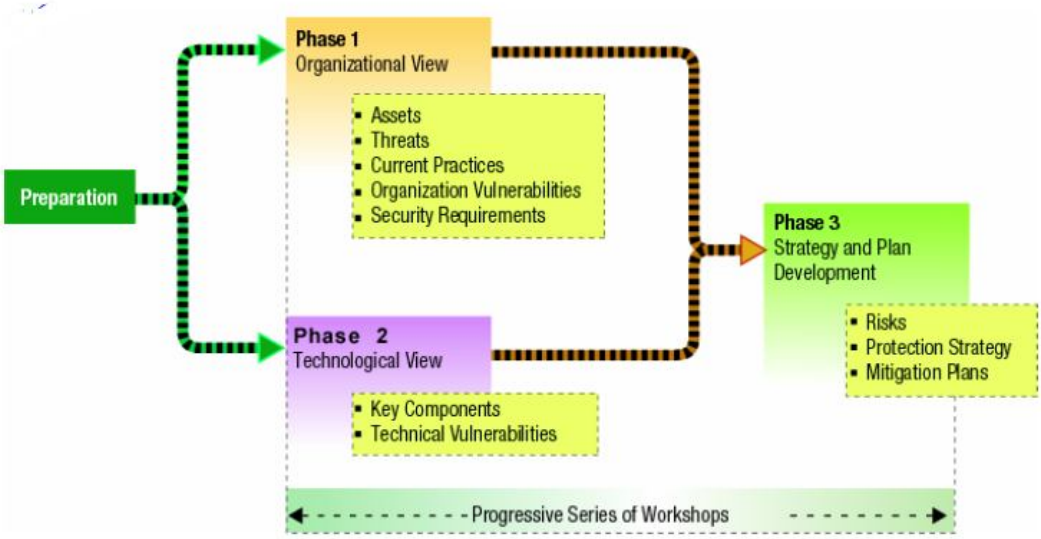

Fig 3.3 OCTAVE phases (adapted from [61])

The COSO ERM Board has confirmed that the key concepts and principals embedded in the original frame work remain essentially, besides that the updated framework develops principles and supporting points of focus within each of the five foundational components of internal control; control environment, risk assessment, control activities, information and communication, monitoring activities. The COSO ERM new framework involves two rounds of public exposure to review, refresh, and modernize the original framework, ensuring it remains relevant [62]. COSO's Internal Control- Integrated Framework enables organizations to develop systems of internal control that adapted to changing business and operating environment, mitigate risks to acceptable levels, and support organization's decision makers. The framework has three categories of objectives, which are what an entity strives to achieve, these objectives are: operations objectives, which concern to effectiveness and efficiency of the entity's operations, reporting objectives. These are concerned to internal and external financial and non-financial reporting, and compliance objectives which concern to obligate to laws and regulations to which entity is subject.

Internal control consists of five components, control environment, risk assessment, control activities, information and communication, and monitoring activities, they are represented what is required to achieve the objectives. The control environment is the

collection of procedure, structure, and criterions that supply the foundation for implementing internal control across the organization, risk assessment which involve a repeated and dynamic process for identifying and assessing risks to the accomplishment of objectives, establishment of the objectives, linked at different levels of the entity form a precondition to risk assessment. Control activities are the actions which performed at all levels and various stages of the within business process, of the entity, and established out of procedures and policies, these actions are established through policies and procedures to ensure that the steps done by the decision makers to mitigate risks are carried out. Information and communication, entity need information to implement internal control responsibilities to support the achievement of its objectives. Communication is repeated process of obtaining necessary information. In the last component monitoring activities, two kinds of evaluations, ongoing and separate are used separately or some combination of them to check whether each of five components of internal control is present and functioning. Results are evaluated against standard affirmed by regulators. The relationship between objectives, component, and organizational structure depicted in the form of cube [64]. Figure 3.4 represents this relationship.

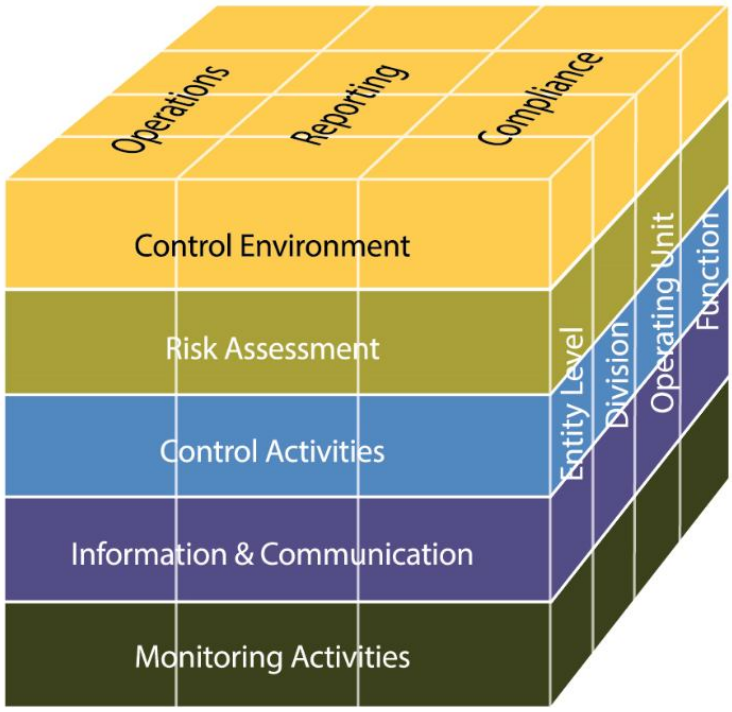

Fig 3.4 The COSO ERM Cube (adapted from [62])

The framework sets out seventeen principles representing the fundamental concepts associated with each component. An entity can achieve effective internal control by applying all principles. In spite of internal control provides reasonable assurance of achieving the entity's objectives, some limitation may exist because internal control cannot prevent bad decisions, or external events that can lead to failure to achieve the organizational operational goals.

International Organization for Standardization (ISO) standard 31000 [63] was published in 2009 as an internationally agreed standard for the implementation of risk management principles. “Risk Management- Principles and guidelines” is the title of ISO 31000 standard. The standard provide vision, guidance, and generic iterative process of risk assessment in organization of any size, and it can be applied to any type of risk regard less of its nature or effect type (positive or negative) [63, 65]. The risk management process is governed by the principles, which establish the values and philosophy of the process. Risk management principles link the framework and practice of risk management to the strategic goals of the entity. The principles help align risk management to corporate activities, besides supporting a comprehensive and coordinated view of risk that implemented to specific organization [58]. The risk lifecycle framework in ISO 3100 consists of 5 components, mandate and commitment, design of framework, implement risk management, monitor and review framework, and improve framework. Figure 3 represents the ISO 31000 risk management process, the key stages are: risk assessment and risk treatment. The risk assessment stage consists of risk identification, risk analysis, and risk evaluation. Risk identification sets up the discovery of the organization to risk and uncertainty. The outcome of risk analysis can be used to make a risk profile, which can be used to rank the relative importance of each identified risk. This process allows the risks to be mapped to the business area affected. Risk evaluation prioritized risk control action in terms of their potential to benefit the organization. The risk treatment stage is defined in ISO 31000 as the activity of selecting and implementing appropriate control measures to modify the risk. It include risk control (mitigation), risk avoidance, risk transfer, and risk financing. ISO 31000 is not intended for the purpose of certification [63, 65].

Technical committee ISO/TC 176 is responsible for the ISO 9000 family series of standard for quality management and quality assurance. It was first published in 1987 [66, 67]. The series provide guidance and tools for companies and organizations who want to ensure that their products and service consistently meet customer's requirements, and that quality is consistently improved [67]. ISO 9001: 2008 proposed approach that required improving the processes operating in the organization. Risk management has no place in this approach [65]. The purpose of ISO 9001 is to provide organizations with a foundation upon which to build sound business practices and processes [67]. Risk management is strongly suggested by the ISO 9004: 2009 under the title "managing for the sustained success of an organization - A quality management approach"[65]. The purpose of the developed of ISO 9004 was to maintain consistency with ISO 9001, and to help organizations who are users of ISO 9001 obtain long term benefit from the implementation of a more broad- based and in- depth impact quality management system [68, 69]. Figure 3.5 shows ISO 31000 risk management process.

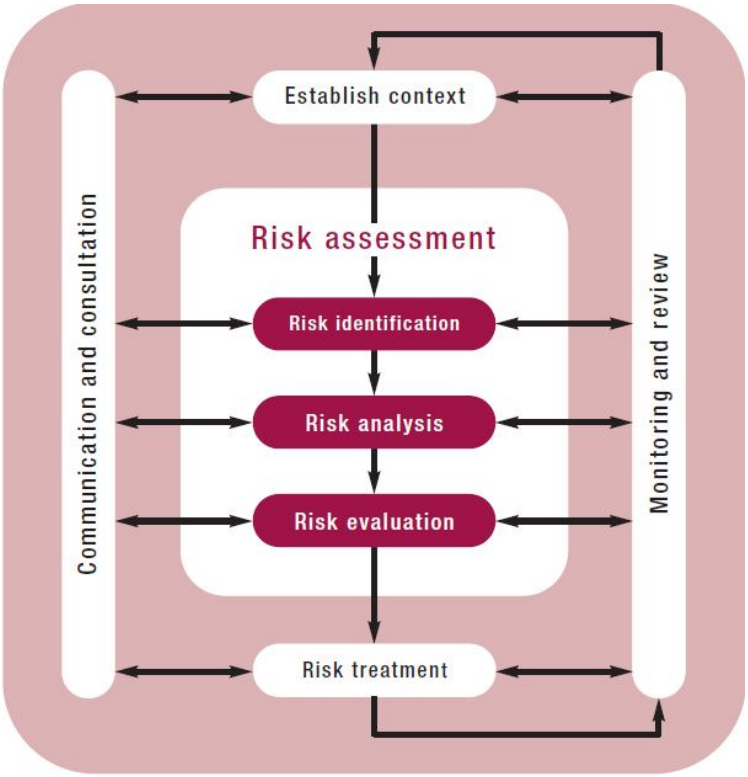

Fig 3.5 Risk management process based on ISO 31000 (adapted from [63])

Control Objectives for Information and Technology framework (COBIT) was introduced in 1996 by ISACA (Information Systems Audit & Control Association) and IT Governance Institute [70-72]. COBIT is a comprehensive framework for information technology governance ITG that helps enterprises to create optimal value from IT by maintaining a balance between realizing benefits and optimizing risk levels and resource use [70, 72, 73]. The purpose of COBIT is to define a series of process necessary for steering IT resources to achieve business objectives. The framework is based on assessable controls and guided by a Capability Maturity Model (CMM) to facilitate the identification of risk exposures and realization of benefit, thus, it is used as a compliance checking system [72]. COBIT is not specific to information technology, it addresses information technology governance, and refers to information security among many other issues, it divides the information technology governance into 34 processes, and provides a control objective for each of these 34 process [70, 71].

The last edition of the framework, COBIT 5, published in 2012 was introduced for "Enterprise governance of IT" [74]. It saw a shift in the framework's orientation towards business through integrate all ISACA models (such as Val IT, Risk IT) into one integrated model. Additionally, it separates governance from management and focuses on board-level concerns [72, 75]. COBIT 5 identifies five basic principles, and seven category of enablers shown in Figures 3.6, and 3.7 respectively, to govern and manage the information requirements [70, 71]. COBIT 5 also introduces a new process-reference model, new processes, update and expanded goals and metrics, and alignment with ISO/IEC 15504 process capability-assessment model [75].

Microsoft developed cloud risk decision framework [76], that was based on ISO 31000 standard. It serve as support in decision-making process as per risk management best practice guidance outlined in ISO 31000. This guidance was designed to help the organization to objectively identify, analyze, assess, and determine potential risk treatment alternatives for risks related to cloud strategy which organization planned to adopt it.

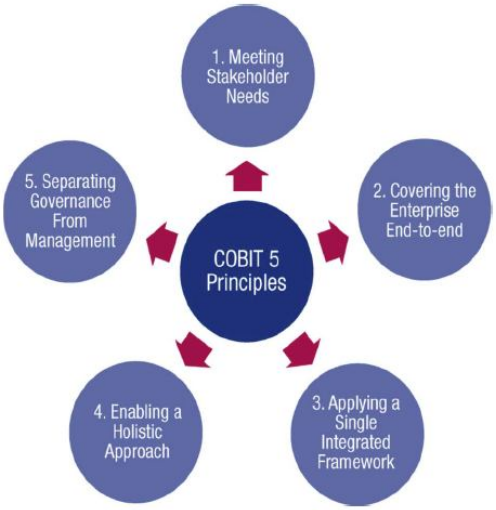

Fig 3.6 COBIT 5 principles (adapted from [75])

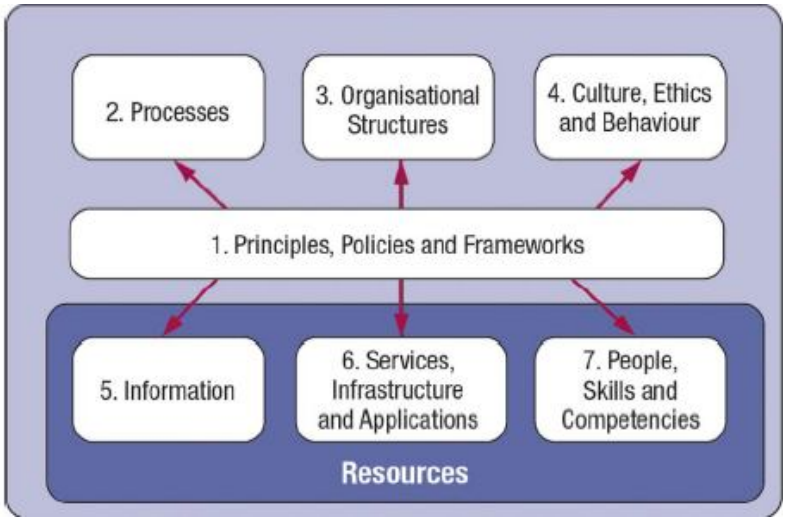

Fig 3.7 COBIT 5 enablers (adapted from [75])

Despite all best practices and recommendations, the experience of practitioners shows that there is little evidence that risk management is being efficiently applied in a systematic and periodic way. Actually, most standards guides only provide high level guidelines for general purpose risk assessment in a textual description form. Very low information is given about how actually implement these standards in practice and most of the proposed processes are assumed to be manual [77]. The authors of [78-80] have investigated the actual benefits and shortcomings of different approaches for risk management in real life environment. Many issues were arises from these investigation

such as little knowledge reuse, inadequate documentation, and lack of tools to automate, monitor, report, and support decision-making.

Shamala et al. [12] suggested a conceptual framework of info-structure for information security risk assessment. Six methodologies, which are currently available, were compared and analyzed to develop the framework. The aim of the framework is to explain the general view of flow, types of information to be gathered and requirements which need to be met before conducted any risk assessment. A quantitative approach for assessing and evaluating risks was suggested in [81]. This model uses empirical data that reflects the security posture of each vulnerability to calculate loss expectancy; a risk impact estimator. CORAS, was presented in [82], it is a model-based risk management process. The main objective of the CORAS is to develop a framework to support risk assessment of security critical systems.

3.4.2 Risk Assessment models Specific to Cloud Computing

In recent years, the principles and practices of risk assessment were presented to the world of cloud computing either as general methodology or focus on specific type of risk such as SLA fulfillment. The following part discus different research that has been done in the areas of risk assessment in cloud computing environment.

The main research topic of several groups and organizations currently is cloud computing standards. Cloud standards coordination, was formed in July 2009, its main goal “is to create a landscape of cloud standards work, including common terminology” [31]. They created a wiki page [83], where different cloud oriented Standard Developing Organizations (SDOs) can update their parts of research. Here is a brief description about each of SDOs research areas.

Cloud Standard Customer Council (CSCC) [84], is an end user advocacy group dedicated to accelerating cloud's successful adoption, and drilling down into the standards, security and interoperability issues surrounding the transition to the cloud. CSCC provides cloud users with the opportunity to drive client requirements into standards development

organizations and deliver materials such as best practices and use cases to assist other enterprises.

In its recommendations on risk assessment for cloud computing, European Network and Information Security Agency (ENISA) [36], investigated the different security risks related to cloud computing. It provides a list of relevant incidents scenarios, vulnerability and assets. ENISA estimate the level of risk on the basis of likelihood of a risk scenario mapped against the estimated negative impact. National Institute of Standards and Technology (NIST) [85], develop a guidelines to use by the organizations that process sensitive information, such as federal organizations; also non-governmental organizations can use this guidelines. The purpose of the NIST guideline is to provide a foundation for the development of an effective risk management program. The primary goal of this document is to help organizations to better manage IT related risks. The risk assessment methodology encompasses nine primary steps: system characterization, threat identification, vulnerability identification, control analysis, likelihood determination, impact analysis, likelihood determination, impact analysis, risk determination, control recommendation, and results documentation

Cloud Security Alliance (CSA) is a non-profit organization promoting the use of best practices, common level of understanding, awareness and guidelines for cloud related security threats. In December 2009, CSA released “Security Guidance For Critical Areas of Focus in Cloud Computing” [86], where they identified thirteen areas of concerns in three major sections. This document quickly becomes the industry-standard catalogue of best practices to secure cloud computing, therefore many business, organization, and governments have incorporated this guidance into their cloud strategies. Security, Trust & Assurance Registry STAR [87] is a powerful program for security assurance in the cloud, provided by Cloud Security Alliance CSA. STAR program includes a complimentary registry that that documents the security controls provided by popular cloud computing offering. CSA STAR based upon two key research component: CSA Cloud Control Matrix CCM [88], CSA Consensus Assessment Initiative Questionnaire CAIQ [89]. Cloud Control Matrix CCM, is designed to provide essential security principles to guide cloud providers and to help cloud customers in assessing the overall

security risk of a cloud provider. CCM provides organizations with the needed structure, detail and clarity relating to information security tailored to the cloud industry. Consensus Assessment Initiative Questionnaire CAIQ is a set of “yes”, “no” questions a cloud customer may hope to ask of a cloud provider. The questions are based on security controls found in the CSA Cloud Control Matrix.

The design, implementation of an effective and efficient risk assessment framework for cloud service provision associated with corresponding mitigation strategies is proposed in [51]. The framework aim to analyze and address the risk factor in a cloud service ecosystem and it provides technological assurance that will lead to higher confidence in cloud providers together with cost effective, reliable and productive cloud service provider's resources. It emphasizes that risk must be considered at each service stage in relation to the assets which need to be protected besides it must be performs at service providers (SP) and infrastructure providers (IP) levels. Service provider need to identify risks during the service deployment and operation, SP needs to know and assessed risks of each IP, this enable them to meet their responsibility about matching the end user requirements with the correct IP. Infrastructure provider performs risk assessment during admission control and internal operations, which increase the performance and quality of the IP. The information about vulnerability, threats, and risks associated with each asset is available in simple database called risk inventory, it developed to determine how certain risks can be managed and evaluated to be brought to an acceptable level. There are various risk models, which can be introduced to choose relevant mitigation strategies so the proposed model could be built as a combination of probabilistic, possibility, and hybrid models and assess risk based of four categories of risks namely technical, policy, general, and legal. Final stage is to implement risk mitigation strategy, with the context of this model the main strategies to be applied are avoidance and limitation. This framework and its software toolkit implementation is part of the research and development work of the OPTIMIS (Optimized Infrastructure Service) project [90].

Anew risk assessment framework for cloud service provision [91] is proposed to assess and improve the reliability and productivity of fulfilling an SLA in a cloud environment. The aim of this model is to allow individuals to negotiate and consume cloud resources

using service level agreement (SLA). The model claims that it is essential to identify what data is required for such risk assessment and how it is going to be analyzed to estimate the actual risk, a risk inventory is developed and used for this purpose. A quantitative risk assessment approach is then applied to measure the level of risk attached to each asset, model methodology divided in to six stages, in risk inventory stage the requirement analysis is carry out to identify how the risk inventory is populated. in the vulnerability, threat identification stage respectively, each vulnerability and threat is represented as a single bit in the vector of theirs; to indicate its existence it gives the value 1 otherwise 0. Data requirement that need supported is identified in data monitoring stage. Simultaneously, in the event analysis as the possibility of an event occurring identified, the likelihood should be estimated. The likelihood of threat acting over vulnerability is defined as $Lji = (Tj, Vi)$. The quantitative risk assessment approach is applied at the quantitative risk analysis stage, to estimate the level of risk for each asset. The last stage is model, in which the individual risks are first calculated, and then to enhance knowledge an aggregated risk is estimated.

A quantitative Impact and Risk Assessment Framework (QUIRC) [92] presented for analyzing and assessing the risks and impacts to the security of cloud-based software deployment. This framework categorize risks based on security objectives (SO), which are defined base on the potential impact on an organization. Three of these objectives confidentiality, integrity, and availability; are defined by Federal Information Security Management FISMA [93] as they are addressed in the context of traditional network and systems security. In the context cloud plat forms, [94] defined multi-party trust and mutual auditability. Usability objective is added by QUIRC authors as one requirement unique to cloud computing platforms. These security objectives can be referred as CIAMAU framework. The conceptual basis of this framework is come from Federal Information Processing Standards FIPS approach, which represent risk as a product of the probability (P_e), of threat which is a fraction less than 1, and its potential impact (I_e) it can assigned to a value on a numerical scale: $Re = P_e I_e$ the framework adopt wide-band Delphi method [95] for evaluation the risk impact based on expert opinion. The method is a forecasting technique used to collect information for assessing risk. Features of the QUIRC methodology is that it gives the vendors, customers and regulation agencies the

ability to comparatively assess the relative robustness offers introduced by different cloud venders. Other feature of QUIRC is that it can help to deal efficiently and relieve the major FUD (fear, uncertainty and doubt). However, its limitation is the meticulous collection of historical data for threat event probability calculation, which requires data input from those to be assessed cloud computing platforms and their vendors [92].

A Semi-Quantitative BLO-driven cloud risk assessment (SEBCRA) [96], presented as core sub-process of a cloud risk management approach. The cloud risk management and its core sub-process allow cloud organization to be aware of cloud risks and align their low-level management decisions according to high-level objectives. This framework is designed to address impacts and consequences of cloud specific risks into BLOs of a given cloud organization, also, it aims to increase the probability of success which lead to decrease both the opportunity to failure and the uncertainty in achieving those objectives. The core process of BLO-driven cloud risk management is Risk Level Estimation (RLEs) as outputs, which are individually specified for each risk and BLO affected. SEBCRA main intent is to rank cloud risks. The assessment methodology of this framework is subdivided in to risk analysis and risk evaluation. In the risk analysis step a proposed semi-quantitative risk analysis which uses a standard risk level matrix in order to show risk level rates (based on ISO/IEC 27005:2008). The risk analysis step divided to three stages: risk identification, risk description, risk estimation. The organization's potential risks are defined in risk identification stage; A comprehensive risk assessment method guarantees at risk description stage; and in risk estimation stage the likelihood of occurrence and the estimated impact on BLOs of each defined risk is figured out. To evaluate the impact, authors use 10x5 risk level matrix because they consider five possibilities for either positive or negative impacts. After assessing risk, the sub-process Risk Treatment is used to define potential risk-aware actions, controls, and policies to decide which Risk Mitigation methodology (avoid, reduce, accept, and transfer) which aims to move risks on the negligible or profitable levels.

Morali et al. [97] introduced (CARC++), an extended version of Confidentiality Risk assessment and Comparison (CRAC) method [98]. CARC++ support decisions about confidentiality requirements. The aim of this method is to enable the specification of

confidentiality requirements in an SLA between a client and IT resource provider. The method elucidate that to do confidentiality level specification must be chosen, the method chosen must satisfy at least three criteria: the confidentiality levels must not be specify as percentages of data loss; it is not of observing episodes; and it does not require a provider to uncover confidential information. The Confidentiality Risk Assessment and Comparison (CRAC) [98] provides confidentiality risks of two alternative networked IT architectures by analyzing how information can flow through a network, and how unauthorized individuals can move through the network.

CRAC++ expand CRAC with a step to recognize confidentiality requirements of the customers that are not implicitly by the known confidentiality requirements of the provider, and which are candidates for inclusion in an SLA with that provider. CRAC++ consist of four steps: at the end of step 0, the risk assessor is provided with the data which is used in the next steps of the method, these data is: information assets, IT architectural components, threat agents, relevant vulnerabilities and confidentiality. In the step of assessing total impact of disclosure per component, an information flow is made for each information asset and each component that the asset can reside on, at the end, the components for which unauthorized access would create a total impact higher than a certain value is identified and it determined by system owners. The likelihood that a component will be accessed by an unauthorized agent is assessed in step two, assessing protection level per component. Third step: determining candidate confidentiality requirements, the confidentiality of requirements of the client that are not implied by known confidentiality requirements of the provider are identified. This step is subdivided in to three stages, first they identify vulnerabilities against which the client wants to protect itself. In second stage the protection levels under the assumption that the clients confidentiality requirements were satisfied were identified. Finally, the comparison of the protection levels of critical components in the best and worst cases is done by confidentiality expert, besides that the expert identifies the confidentiality requirements that the provider must satisfy.

An information security risk management framework for cloud computing environments was presented in [99]. The purpose of this framework is to identify threat, and

vulnerability, and for better understanding critical areas in cloud environment. The framework can be applied to all cloud computing service and deployment models. This framework was developed based on evolving ISO/IEC 27001 standards [100], NIST risk management guide for information technology systems [101], and Booz Allen Hamilton information security governance government consideration for the cloud computing environment [102].

The framework consists of seven processes that is embedded in three phases. First phase is: Architecture and establishing the risk management program (PLAN) and it consists of two processes; selecting relevant critical area, and strategy and planning. The second phase implements and operate encompasses risk analysis, risk assessment, and risk mitigation. Monitoring and review is the last phase and it include the process of assessing and monitoring, and risk management review processes [99].

Cloud Adoption Risk Assessment Model (CARAM) framework designed by [103]. CARAM is a qualitative risk assessment model designed for helping cloud customers to assess risks that they face by selecting a specific cloud provider. CARAM is based on existing frameworks such as ENISA, CSA, CNIL, and CAIQ and complements them to provide the cloud service customer with a practical tool. The limitation of this method that its accuracy of risk assessment depends on the accuracy of the input data and the appropriateness of the proposed formulas.

A quantitative security assessment approach for cloud Security Level Agreements is proposed in [41]. This approach utilize the Reference Evaluation Methodology (REM), originally proposed in [104] as a technique to quantitatively evaluate security policies. A novel security risk assessment model for information system in cloud computing developed by [104]. This model is based on Analytical Hierarchy Process AHP, AHP is applied for optimal decision-making and to achieve weighting factor. This work also summarized 8 kinds of threats to security principles, and lists the corresponding factors. Combing with collaborative and virtualization of cloud computing technology and so on, adopting the theory of AHP and introducing the correlation coefficient to analyze the multiple objective decision.

A cloud-based assessment as a service paradigm is proposed [105] as a promising alternative. A framework called SecAgreement (SecAg) [106], that extends the current SLA negotiation standard, WS-Agreement [107], to enable the description of security metrics on service description terms and service level objectives of the SLA. The framework allow organizations to quantify risk, identify any policy compliance gaps that might exist, and as a result select the cloud services that best meet their security needs. Bernsmed et al. [42] outlines a framework for security SLAs for federated cloud services, in the context of hybrid clouds. The purpose of this method is twofold: to facilitate rapid service composition and agreements based on the necessary security requirements, and to establish trust between the customer and the providers. Carrol et al. [14] provides recommendations for the mitigation of cloud computing security risks as a fundamental step towards the development of guidelines and standards for secure cloud computing environments.

Luna et al. [108] introduced the basic building blocks of a proposed security metrics framework for cloud provider's security regarding to the different service and deployment models of the cloud. The framework targets to improve tasks such as dependability assessment or compliance evaluation. The author's goal is to create an open, flexible and technology-a gnostic framework able to be extended through the integration of new security metrics. Lenkala et al. [109] presented a risk assessment framework to study the security risk of the cloud carrier between cloud users and two cloud providers. This framework enables cloud users to select quality of security services among cloud providers, by providing the quantifiable security metrics of each cloud carrier. Wang et al. [110] designed a method of the cloud computing security management risk assessment. It mainly commits to assessing the Cloud Computing Security Management Risks (CCSMR) to clarify distribution of the risks, occurrence possibilities, correlation between risks and assets, impact level, correlation between risks and vulnerabilities.

A case study for cloud computing risk assessment was presented in [111], it represents a one-time attempt at risk assessment of the cloud computing arrangement. Xie et al. [112] analyzed the characteristic of personal cloud computing, and built an industry chain

frame work of personal cloud computing, which is based on the current cloud industry chain in China. Collaboration-Based Cloud Computing Security Management Framework for cloud computing was introduced in [113]. The framework based on aligning the FISMA (Federal Information Security Management) standard, to fit with the cloud computing model, which enable cloud providers and consumers to be security certified. The framework goal is to improve the collaboration between cloud providers, service providers, and service consumers in managing the security of the cloud platform and the hosted services.

Authors in [114, 115] presented a cloud-based service security lab, to specify security requirements on web services and cloud web applications composed of web services. This cloud platform enables the testing, monitoring and analysis of Web Services regarding different security configurations, concepts and infrastructure components. Bertram et al. [116] proposed a novel service-oriented infrastructure (SOI), which aims to provide a platform-as-a-service (PaaS) solution for on-demand management of security risks associated with assets shared in clouds. The authors assumed a trusted and secured cloud platform with a focus to provide security PaaS that can manage and mitigate security risks of the services shared between two collaborating enterprises.

3.4.3 Risk Assessment Models using Machine Learning Techniques

Risk assessment has been discussed by many researches in different areas, and the opinions about risk and the association of its dependent variables is differ. Soft computing and machine learning tools provide an excellent framework to model risk. In this section, we will present risk assessment models and the use of machine learning tools to develop risk assessment models in conventional systems.

Haslum et al. [117] presented the implementation of the Hierarchical Neuro-Fuzzy online Risk Assessment (HiNFRA) model in intrusion detection systems. This model used a fuzzy logic approach, the fine tuning of fuzzy logic is achieved using neural network learning techniques. to develop the model. Further, the authors. Abraham et al. [118] proposed the use of Genetic Programming approach for risk assessment in an Intrusion Detection System (IDS). Yucel et al. [119] developed a predictive risk assessment model

for a hospital information system HIS to estimate risk before the implementation of new HIS.

3.5 Problems with Existing Risk Assessment Methodologies

1- Some problems associated with risk assessment methodologies that are based on being able to accurately quantify reliability. It cannot be considered as a good approach because 100% reliability does not necessarily correlate to zero percent risk; the existence reliability models are conflicting and make it almost impossible to know the true reliability of a part of software.
2- Risk/threat identification models and tools provide a starting point in dealing with risk because they only provides list of risk/threat identification. In this case not all risk identified can be addressed.
3- The risk impact estimation approach, their problem is how to know what asset costs to use. It also suffers from the lack of sufficient data to determine probability of loss [81].
4- Most of existing risk assessment models which they are specific to cloud computing environment are theoretical models and they didn't implement machine learning techniques to assess risks.

3.6 Summary

This chapter presents the definition of risk, risk management, risk assessment and it's different stages. In addition, the Chapter illustrated the existing generic risk assessment methods besides risk assessment specific to cloud computing. Towards the end the problems of existing risk assessment approaches is also provided.

Chapter Four

Machine Learning Techniques

As the information technology world grows more complex, the amount of data therein, in our lives, seems to increase, and there is no end in sight. Knowledge discovery provides development of methods and techniques for making sense of data. The basic problem addressed by the knowledge discovery process is one of mapping low-level data into other forms that might be more compact, more abstractor more useful. While knowledge discovery refers to the overall process of discovering useful knowledge from data, data mining refers to a particular step in this process [120].

4.1 Data Mining

The manual process of data analysis becomes much more tedious as the size of data grows and the number of dimensions increases. For this reason the process of data analysis needs to be computerized. Data Mining is an iterative process within which the progress is defined by discovery of earlier, unidentified, valid patterns and relationship in large dataset, through either automatic or manual tools [121-123]. Data mining becomes the only hope to find the regularities deeply buried in the data [124, 125].

Machine learning provides the technical basis of data mining [125]. Data mining is a machine learning discipline and, is inspired by pattern recognition. Many researchers try to define data mining. [126] defined data mining as “the process of using variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions”. Another data mining definition presented by [125]: “the process of discovering patterns in data”. This process can be automatic or semiautomatic. The discovered pattern must be meaningful and entails some advantage. Data mining is a particular step in Knowledge discovery (KD) process It is the application of specific algorithms for extracting of patterns from data [120]. Data mining involves the use of

sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data set that allow the prediction of future results. These tools can include modeling techniques, statistical analysis, database technology and machine learning. Data mining finds patterns and/or relationships in data and infers rules [121, 127].

4.1.1 Data Mining Methods

Data mining is the extraction of implicit, previously unknown, and potentially useful information from data [125]. The knowledge extracted from data mining allow the user to find and tune interesting patterns in the data to help in the process of decision making [124]. The tasks of data mining can be classified into two main categories: “descriptive” and “predictive”. The descriptive characterizes the general properties of the data in the database to finds the patterns for presentation to a user in a human-understandable form. On the other hand, the predictive provides for creation of models that are capable of producing prediction results when applied to unseen data [120, 124]. There are a variety of particular data-mining methods that can be used to achieve the goals of prediction and description tasks:

Classification: is a learning function that maps a data item into one of several predefined classes. It can be considered as a supervised technique where each instances belongs to a class [124]. Classification process aims to build a model from classified objects in order to classify previously unseen objects as accurately as possible [123].

Regression: Regression is a data mining technique used to fit an equation to a dataset. It is a learning function that maps a data item to a real valued prediction variable. The simple form of regression can be represent by ($y = mx + b$). Multiple regression allows the use of more than one input variable and allows for fitting of more complex models such as quadratic equation [120].

Clustering: is a descriptive task, which identifies a finite set of categories or clusters to describe the data. These categories can be mutually exclusive and exhaustive or contain of a richer representation [120] .

4.1.2 Data Mining Learning Approaches

Data mining is the application of specific algorithms for extracting patterns from data. The data mining algorithms can follow three different learning approaches: supervised, unsupervised or semi-supervised [124].

Supervised learning: In this approach the algorithm works with a set of examples of known labels. In other words, the algorithm is given the desired outputs and its goal is to learn to produce the correct output given a new input. The labels can be nominal values in the case of classification task or numerical values in the case of the regression task.

Unsupervised learning: In contrast, the labels of the examples in the dataset are unknown. The algorithm provides for sorting of examples according to the similarity of their attribute values.

Semi-supervised learning: This approach is used when a small subset of labeled examples is available together with a large number of unlabeled examples.

4.1.3 Preprocessing Stages

Many factors affect the success of machine learning algorithm on a given task. The representation and quality of the example data is the first and foremost. Data available for mining is raw data It may be in different formats and comes from different sources It may consist of noisy data, irrelevant attributes, missing data etc. Data needs to be preprocessed before applying any kind of data mining algorithm, a process which is done using the following steps [124, 128]:

Data integration: this step is done when the data comes from several different sources. In this case the data needs to be integrated which involves removing of inconsistencies in names of attributes and/or attribute values.

Data cleaning: It involves detecting and correcting errors in the data.

Discretization: It applies when the data mining algorithm cannot cope with continuous attributes. In this step the continuous attributes are transforming into categorical attributes. Discretization often improves the comprehensibility of the discovered knowledge.

Dimension reduction: Dimension reduction methods are usually based on mathematical projections, which attempt to transform the original features into an appropriate feature space. After dimension reduction, the original meaning of the features is usually lost [129].

Feature selection: Not all attributes are relevant, so in this step a subset of relevant attributes is selected for mining. In other words feature selection methods directly select some original features to use, and therefore they can preserve the original meaning of features, a very desirable quality in many applications.

4.2 Feature selection

Many factors affect the success of machine learning algorithm for a given task. The representation and quality of the training data is first and foremost. Theoretically, having more features should result in more discriminating power. However, practical experience with machine learning algorithms has shown that this is not always the case. Many learning algorithms can be viewed as making (biased) estimate of the probability of the class label given a set of features. This is a complex, high dimensional distribution. If there is too much irrelevant and redundant information present or the data is noisy and unreliable, then learning during the training phase is more difficult. Variable and feature selection have become the focus of much research in areas of applications for which datasets with tens or hundreds of thousands of variables are available. Many irrelevant attributes may be present in data to be mined. So they need to be removed. Moreover, many mining algorithms don't perform well with large amount of features or attributes [124].

Feature selection is the process of extracting subset instances from the original data set. It presents an important technique in data preprocessing in data mining [130] This reduces the dimensionality of the data and enables data mining algorithms to operate faster and more effectively. In some cases, accuracy on future classification can be improved; in others, the result is a more compact, easily interpreted representation of the target concept. Feature selection techniques needs to be applied before any kind of mining algorithm is applied. The main objectives of feature selection are: to avoid

overfitting and improve model performance, to provide faster and more cost-effective models, and to provide a better understanding of the underlying process that generated the data [131-133].

Attribute selection reduces dataset size by removing irrelevant and redundant attributes. Appling feature selection technique involves both search algorithm and an evaluation algorithm. Feature selection algorithm generates and compares possible solution to proposed subset of features and attempts to find an optimal subset. In order to perform this task it needs to address basic issues that affect the nature of search [134]:

1- Starting point: It determines the search direction; the algorithm can begin with no features and successively add attributes. In this case, the search is said to proceed forward through space. Conversely, the search may proceed backward through the search space. In this case, the search starts with all features and successively removes them. Third option is to begin somewhere in the middle and move out wards from this point.

2- Search organization: A heuristic search can give good results. An exhaustive search is prohibitive for all but small initial numbers of features. Both, heuristic and exhaustive they do not guarantee finding the optimal subset.

3- Evaluation strategy: A single important factor to differentiate among feature selection algorithms is the process of evaluating feature subset. One way, called the filter, operates independent of any learning algorithm Undesirable features are filtered out of the data before learning starts. The other method dubbed wrapper uses an induction algorithm along with a statistical resampling technique such as cross-validation to estimate the final accuracy of feature subsets.

4- Stopping criterion: The search algorithm must apply a specific criterion to decide when to stop searching through the space of feature subsets. Depending on the evaluation strategy, the search algorithm might stop adding or removing features when none of the alternatives improves upon merit of a current feature selection. Alternatively, the algorithm might continue to revise the feature subset as long as the merit does not degrade. A further option could be to continue generating feature subsets until reaching the opposite end of the search space and then select the best.

Applying feature selection techniques involves an extra layer of complexity. Instead of finding optimal parameters for full set of features; we need to find the optimal subset feature first [124, 135]. There are many potential benefits of feature selection such as: reducing the measurement and storage requirements, facilitating data visualization and data understanding, reducing training and utilization times, and defying the curse of dimensionality to improve prediction performance [131].

Feature selection process is divided broadly into two approaches: *filter approach*, and *wrapper approach,* based on their dependence on the inductive algorithm that will finally use the selected subset [124, 132, 136-138].

Filter approach: In the filter approach, the attribute selection methods operate independently of any data mining algorithm, undesirable features are filtered out of the data before induction commences. The subset of features left is presented as input to the data mining algorithm. The advantages of filter techniques include simple and fast computation and easily scalable for high-dimensional datasets and it needs to be performed only once because it independent of the mining algorithm. Its disadvantage is that filter approach ignores the feature dependencies, which may lead to worse classification performance when compared to other types of feature selection techniques.

Wrapper approach*:* Wrapper methods have borrowed search and evaluation techniques from statistics and pattern recognition. In this approach the feature selection method uses the result of data mining algorithm to estimate the accuracy of feature subsets via statistical re-sampling technique. The major characteristics of the wrapper approach is that the quality of an attribute subset is directly measured by the performance of the data mining algorithm applied to that attribute subset, and the ability to take into account feature dependencies. The common drawbacks of wrapper approach include a higher risk of overfitting than filter approach and it is computationally intensive. Filter approach has proven to be much faster than wrapper and hence can be applied to large data sets containing many features. Another method [139] was introduced, termed embedded technique, in which search for an optimal subset of features is built into the classifier construction, and can be seen as a search in the combined space of feature subset and hypotheses. Termed embedded just like wrapper, is specific to a given learning

algorithm. It has the advantage of interaction with the classification model, while at the same time being far less computationally intensive than wrapper approach.

4.3 Machine Learning Techniques used to Build the Model

In this section the methods used to develop the risk assessment model are discussed:

4.3.1 Decision Trees

Trees classify instances by sorting them based on feature value This process starts at root node. Each node in a decision tree represents a feature and each branch represents a value that the node can assume. DT is developed through an iterative process of splitting data into discreet groups. The feature that best splits the data would be the root node of the tree. How the split is done depends on algorithm used to implement [121, 140]. There are numerous methods for finding the feature that best divides the data. However, a majority of studies have concluded that there is no single best method. The same procedure is then repeated on each partition of the divided data, creating sub-trees until the training data is divided into subsets of the same class [121]. From a given data set, it is possible to construct as many DTs as possible; some of these trees are more accurate than others. To find the optimal tree is computationally impossible when the search space is large. Efficient algorithms have been developed to induce a reasonable accuracy within a reasonable amount of time. Hunt's algorithm is one of these algorithms, which forms the basis of many existing decision tree induction algorithms. To search the attribute space, these algorithms usually employ greedy strategy in searching the attributes space [125], that constructs decision trees in a top-down recursive divide-and-conquer manner. The advantage of decision tree over other techniques is the output it produces. The output of a decision tree is transparent, which makes it easy for users or non-professional persons to understand [140]. The induction tree algorithm can use two common approaches to avoid overfitting training data: i) Stoop the training algorithm before it reaches a point at which it perfectly fits the training data, ii) Prune the induced decision tree. If the two trees

employ the same kind of tests and have the same prediction accuracy, the one with fewer leaves is usually preferred [121].

4.3.2 Instance-Based Learning (IBL)

Instance-based learning algorithms are classified under statistical methods. They are lazy-learning algorithms, because they delay the induction or generalization process until the regression is performed. Lazy-learning algorithms require less computation time during the training phase than eager-learning algorithms (such as decision trees, neural and bayes nets) but more computation time during the regression process [121, 133].

Instance-based learners searches the patterns space for the k training instances that are closest to the unknown instances, then an instance is classified by comparing it to a data base of pre-classified examples. They follow the assumption that similar instances will have similar classifications. Instance-based learners have three components: distance function which determines how similar two instance are; classification function which specifies how instance similarities yield a final classification for the new instance; a concept description updater which determines whether new instances should be added to the instance database and which instance from the database should be used in classification [141].

Nearest neighbor algorithms are the most straightforward of IBL. They assign equal weight to each instance, then after the instance has been classified it is moved to the instance data base along with correct classification. K-nearest neighbor algorithms are an example of the of IBL complex algorithms. K-nearest neighbor algorithms are based on the principle that the instances within a dataset will generally exist in close proximity properties [133]. K-nearest neighbor algorithms may filter which instances are added to the instance data base to reduce storage requirements and improve tolerance to noisy data [121, 141]. Compared to other algorithms, Instance-based learning algorithms need more time to predict the test samples' classes [142].

4.3.3 Neural Network

Artificial neural network (ANN) is a mathematical model or computational model based on biological neural networks. It is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase [124]. The neural network depends upon three fundamental aspects; input and activation function of the unit, network architecture and the weight of each input connection [133].

To determine input-output mapping, the network first trained on a set of paired data. Then, the weights of the connections between neurons are fixed and the network is used to determine the classification of a new set of data. During a regression the signal at the input units propagates all the way through the net to determine the activation values at all the output neurons. Each input neuron has an activation value that represents some feature external to the net. The activation value is calculated using simple activation function, which sums together the contributions of all sending neurons, where the contribution of neuron is defined as the weight of the connection between the sending and receiving neurons multiplied by the sending neuron's activation value. This sum is then modified unless a threshold level for that sum is reached. Then every input neuron sends its activation value to each of the hidden neurons to which it is connected. The task of these hidden neurons is to calculate its own activation value. Signals are then passed on to output neurons [133].

The determination of hidden neurons size is a problem, because an underestimate of the number of neurons can lead to poor approximation and generalization capabilities, while too much nodes can result in overfitting and eventually make the search for global optimum more difficult [143].

To train the network there are several learning algorithms. The most well-known learning algorithm and widely used to estimate the values of the weights is the back propagation (BP) algorithm. To reach a good weight configuration, back propagation need to perform a number of weight modifications. The greatest problem with feed forward neural networks is that they are too slow for most applications. One approach to speed up the training rate is to estimate optimal initial weights. There are other several

methods for training multilayered feed forward ANN such as: weight-elimination algorithm, genetic algorithms, Bayesian methods. Recently, to improve ANN training algorithms by changing the architecture of network, a number of techniques have emerged These techniques include: pruning and constructive algorithms [133].

4.3.4 Static Regression

Static methods of regression have successfully been applied to functional approximation. Linear regression is one method of static regression; which tries to fit the input-output by linear function. There are two types of linear regression: simple linear regression refers to the regression of y on only one input variable x; multiple linear regression refers to the case where y depends on many input variables x_1, x_2, …. x_n. The projection adjustment by contribution estimation is an extension of linear regression which evaluates the effect of each variable and uses a clustering analysis to give variables various weights of contribution to the linear regression models [144].

4.4 Adaptive Neuro-Fuzzy Inference System (ANFIS)

Knowledge and data that closer to human- like thinking is difficult to represented by system modeling based on mathematical and statistical methods. By contrast a Fuzzy Inference System (FIS) can be viewed as a real-time expert system used to model and utilize human experience, by employing fuzzy if – then rules [145-147].

ANFIS model, which hybridizes an ANN and FIS with a homogeneous structure is used in this research. That is, the ANFIS model integrates the ANN and FIS tools into a compound, meaning that there are no boundaries to differentiate the respective features of ANN and FIS.

4.4.1 ANFIS Architecture

Fuzzy If – Then Rules and FIS: Fuzzy If- then rules are an expression of the form *If **A** Then **B***, where A and B are labeled of fuzzy sets [148] characterized by appropriate membership functions. Fuzzy if- then rules are employed to capture the imprecise modes of reasoning which represent a base role in human decision making [149]. Takagi and Sugeno [150] proposed another form of fuzzy if – then rules, has fuzzy sets involved only in the premise part. The core part of Fuzzy Inference System was represented by fuzzy *If – Then* rules. Fuzzy Inference System is primarily applied to the cases that either if it is difficult to precisely model the system or it is ambiguous to describe the studying issues [151, 152]. The Fuzzy Inference System is the foundation of Adaptive Neuro-Fuzzy System (ANFIS). The drawback of fuzzy logic is that there is no systematic procedure the design of a fuzzy controller. By contrast, a neural network has the ability to learn from the environment, self-organize its structure, and adapt to it in an interactive manner [153].

Adaptive Neuro- Fuzzy Inference System (ANFIS): Adaptive Neuro- Fuzzy Inference System was first introduced by Jang [145]. ANFIS is a multilayer feed forward network, which uses neural network learning algorithms and fuzzy reasoning to map input characteristics into input membership functions (MFs), next, input MFs to a set of if- then rules, rules to a set of output characteristics, then, output characteristics to output MFs, and finally, output MFs to a single-valued output [147].

An ANFIS is, in essence, an ANN that is functionally equivalent to Sugeno first-order fuzzy model. A simple fuzzy inference system rule with two input x and y, and one output z can be expressed as:

Rule 1: if x is A_1 and y is B_1, then

$f_1 = p_1x + q_1y + r_1$ (4.1)

Rule 2: if x is A_2 and y is B_2, then

$f_2 = p_2 x + q_2y + r_2$ (4.2)

Typically, there are six layers in an ANFIS model: one input layer, four hidden layers, and one output layer. Each layer performs a particular task to forward the signals. Such an ANFIS model is shown in Figure 4.1

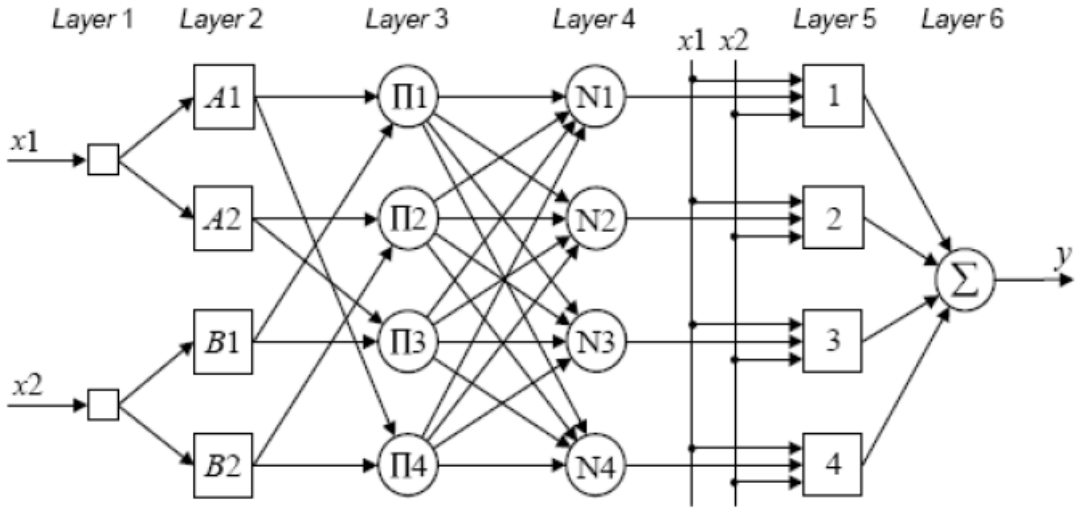

Figure 4.1. ANFIS model architecture with two inputs and one output.

The first layer (i.e. the input layer) of the ANFIS model is the *input layer*. Neurons in this layer simply transmit the external input signals to the next layer.

$$y_i^1 = x_i^1 \qquad (4.3)$$

where x_i^1 is the input signal and y_i^1 is the output signal of neuron *i* in the first layer. The second layer (i.e. the first hidden layer) of the ANFIS model is *fuzzification layer*. Neurons in this layer represent antecedent fuzzy sets of fuzzy rules. A fuzzification neuron here receives an input signal and determines the degree to which this signal belongs to the neuron's fuzzy set. If we let x_i^2 be the input and y_i^2 be the output signal of neuron *i* in the second layer, then we have:

$$y_i^2 = f(x_i^2), \qquad (4.4)$$

where *f* represents the activation function of neuron *i* , and is set to a certain membership function.

The third layer (i.e. the second hidden layer) is the *fuzzy rule layer*. Each neuron in this layer corresponds to a single first-order Sugeno fuzzy rule. A rule neuron receives signals only from the fuzzification neurons which are involved in the antecedents of the fuzzy

rule it represents, and computes the truth value of the rule. In an ANFIS, the 'product' operator is used to evaluate the conjunction of the antecedents [154]. Therefore, we have:

$$y_i^3 = \prod_c^m x_{ci}^3 \quad (4.5)$$

where x_{ci}^3 is the signal from fuzzification neuron *c* in the second layer to neuron *I* in this (i.e. the third) layer; y_i^3 is the output signal of neuron *i* in this layer; and *m* is the number of antecedents of the fuzzy rule neuron *i* represents.

The fourth layer (i.e. the third hidden layer) is the *normalization layer*. Each neuron in this layer receives signals from all rule neurons in the third layer, and calculates the so-called normalized firing strength of a given rule. This strength value represents the contribution of a given rule to the final result [154], and is obtained as:

$$y_i^4 = \frac{x_{di}^4}{\sum_{d=1}^n x_{di}^4} \quad (4.6)$$

where x_{di}^4 4 is the signal from rule neuron *d* in the third layer to neuron *i* in this (i.e. the fourth) layer; y_i^4 is the output signal of neuron *i* in this layer; and *n* is the number of rule neurons in the third layer.

The fifth layer (i.e. the fourth hidden layer) is the *defuzzification layer*. Each neuron in this layer is connected to the respective normalization neuron in the fourth layer, and also receives initial input signals, x_1, x_2,..., x_n. A defuzzification neuron computed the 'weighted consequent value' of a given rule as:

$$y_i^5 = x_i^5 \ (k_{i0} + k_{i1}\, x_1 + k_{i2}\, x_2 + \ldots\ldots + k_{in}\, x_n \quad (4.7)$$

where x_i^5 is the input and y_i^5 is the output signal of neuron *i* in this (i.e. the fifth) layer; and k_{i0}, $k_{i1,}$ k_{i2},..., k_{i2} is a set of consequent parameters of rule *i* [154].

The sixth layer (i.e. the output layer) is the *summation layer*. There is only one neuron in the layer, which calculates the sum of outputs of all defuzzification neurons in the fifth layer, and consequently produces the overall ANFIS output, *y* , as follows:

$$y = \sum_{i=1}^n xi\Sigma \quad (4.8)$$

where *i x* is the signal from defuzzification neuron *i* in the fifth layer to this summation neuron; and *n* is the number of defuzzification neurons, namely the number of fuzzy rules in the ANFIS model.

4.4.2 Training ANFIS Model

Because ANFIS is based on neural network learning, it can be trained to learn from given data. As observed from the ANFIS architecture, in order to construct an ANFIS model for a specific problem, first there is a need to determine the fuzzy rules and the membership functions type. For the fuzzy rules, the antecedent fuzzy sets can be specified according to the problem domain; while for the consequents of the fuzzy rules, the parameters (e.g. $i0\ k$, $i1\ k$, $i2\ k$,..., $in\ k$) are formed and adjusted by certain learning algorithm in the training process. On the other hand, the shapes of membership functions can also be formed and adjusted in the training process.

ANFIS applies a hybrid learning algorithm. This learning algorithm combines the so-called least-squares estimator and the gradient descent method to update the parameters. During the training process, the training dataset is presented to the ANFIS cyclically, and each cycle through all the training examples is called an epoch. In the ANFIS learning algorithm, each epoch comprises of a forward pass and a backward pass. The aim of the forward pass is to form and adjust the consequent parameters, while the purpose of the backward pass is to adjust the parameters of the activation functions.

In the forward pass, when the training dataset is received by the ANFIS model neuron, outputs are calculated on the layer-by-layer basis. The least squares method is used in the forward pass (offline learning) to identify consequent linear parameters, when attempting to minimize the error between the actual state and the desired state of the adaptive network. In the backward pass of ANFIS the gradient descent method is employed in backward pass to tune premise a non-linear parameters, by propagating the error rate from the output end towards the input end, the shape and parameters of the activation functions are updated according to the so-called chain rule [154]

Both consequent parameters and the parameters of activation functions are optimized. The consequent parameters are adjusted and the parameters of activation functions keep fixed in the forward pass, while in the backward pass, the parameters of activation functions are updated and the consequent parameters remain fixed. As a result of the

training process, an optimized model that most fits the training dataset can be obtained, [153-155].

4.5 Ensemble Learning

Both empirical observations and specific machine learning applications confirm that a given learning algorithm outperforms all others for specific problem or for specific subset of the input data, but it is unusual to find a single expert achieving the best results on the overall problem domain. As a result multiple learner systems try to exploit the local different behavior of the base algorithms to enhance the accuracy and the reliability of the overall inductive learning system. An ensemble-based system is obtained by combining diverse models, the base algorithms computed are then collected and combined by another learning process. Therefore, such systems are also known multiple classifier systems, or just ensemble systems. Ensemble learning constitutes one of the main current directions of machine learning research, that are applied to a wide range of real problems. It is mainly used to improve the performance of a model, or reduce the likelihood of an unfortunate selection of a poor one, and to increase the efficiency and accuracy [131, 156].

There are three primary reasons for the use of ensemble learning. The first one is statistical reason, which relates to lack of adequate data to properly represent the data distribution. The second is computational reason which relates to the model selection problem, where among many models that can solve a given problem that we choose. Finally is representational reason that addresses cases when the chosen model cannot properly represent the sought decision boundary. It is important to emphasize that there is no guarantee that the combination of multiple class classifier will always perform better than the best individual classifier in the ensemble [156]. The effectiveness of ensemble methods depends on the accuracy and the diversity of the base learner. An accurate classifier is one that has low error rates. Two classifiers are diverse if they make different errors on new data points [156, 157]. Classifier diversity can be achieved in several ways: i) by using different training datasets to train individual classifiers; ii) by using different

training parameters for different classifiers; iii) by using different type of classifier; iv) by using different features or different subset of existing features [156].

Ensemble learning system consists of two types of learning algorithms: base learner, and combiner. We can distinguish between them as follows: *a base learner* is the result of applying a learning algorithm directly to the row data. *A combiner* is a program generated by a learning algorithm that is trained on the predictions produced by a set of base algorithm on the row data. Both base and combiner is machine learning algorithm [158]. To combine the individual algorithms there are several different combination rules, some of them operate on class labels only, whereas others need continuous outputs that can be interpreted as support given by the classifier to each of the classes. Such rules resemble algebraic combiners, and voting based methods. Algebraic combiners are non-trainable combiners, where continuous valued outputs of classification are combined through an algebraic expression, such as mean, median, minimum, maximum, sum, product. Voting based methods operate on labels only, where $d_{t,j}$ is 1 or 0 depending on whether classifier t chooses j, or not, respectively. The ensemble then chooses class J that receives the largest total vote. Majority voting, and weighted majority voting are two examples for voting based methods [156].

There are many benefits we can get from applying the ensemble learning system. Ensemble learning improves efficiency by executing in parallel the base learning algorithms on subsets of the training data. In addition, it improves predictive performance by combining different learning systems each having different inductive bias, and by combining separately learned concepts, ensemble learning is expected to derive higher level learned model that explains a large database more accurately than any of the individual algorithms [158].

4.6 The Model Performance measurement Methods

Metrics like sensitivity, accuracy, specificity and kappa statistics were used to analyze and compare the performance of machine learning algorithms. Accuracy is the most basic measure of the performance of a learning method. It determines the percentage of

correctly classified instances. Sensitivity gives the percentage of slots in the hypothesis that are correct, whereas specificity gives the percentage of reference slots for which the hypothesis is correct. The kappa statistics is used to measure the agreement between predicted and observed categorization of dataset, while correcting for agreements that occur by chance. In addition to these metrics, the speed of the algorithm, and the time taken to build the model was also considered as a performance indicator [159].

In order to test the effectiveness of our resulting methods, certain standard performance metrics are used in this research. For our problem, we used two statistical measurements: Correlation Coefficient (R), and Root Mean Square Error (RMSE).

$$R = \sqrt{1 - \left(\frac{\sum_{i=1}^{n}(P_i - A_i)^2}{\sum_{i=1}^{n} A_i^2}\right)} \quad (4.9)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(P_i - A_i)^2} \quad (4.10)$$

Where P_i and A_i are actual (desired) and fitted (predicted) output values respectively. So for final result we expect one or near to one value from Correlation Coefficient (CC) metrics, and low values from Root Mean Square Error (RMSE) metrics.

4.7 Summary

This chapter described the methods used in the process of model development. First it introduced data mining methods and learning approaches followed by feature selection. Then, learning methods such as (decision tree, neural network, and static regression) were discussed. Finally, the performance measurement metrics, which are used to evaluate the prediction models, are also presented.

Chapter Five

Research Methodology

This chapter presents the research methodology of the proposed risk assessment model. Fig 5.1illustrated the steps followed to construct the proposed model.

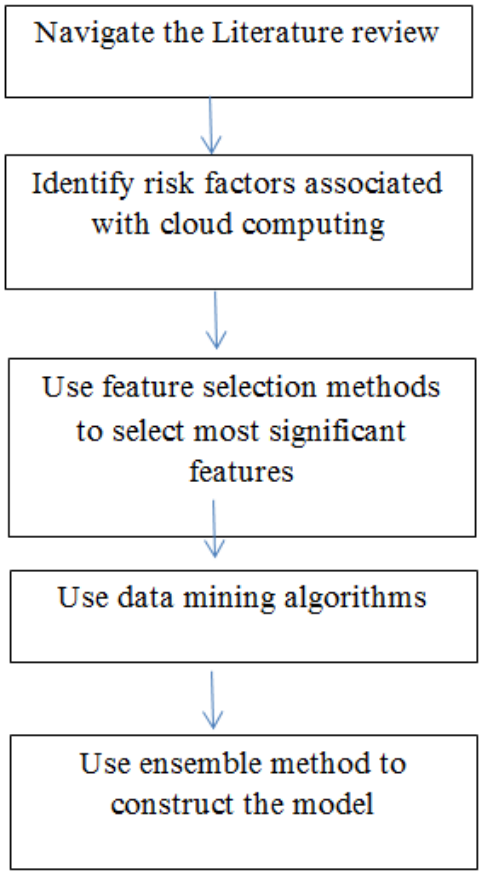

Figure 5.1 the methodology stages used to construct the proposed prediction Model

5.1 Identify Risk Factors and Simulate the Dataset

Cloud computing literature review was navigate and the associated risk factors are founded. Some issues were observed. Many researchers define same risk factor but they used different names. Other researchers define risk factors but it can be merged or

included under another name/category. Thus, with regards to these issues 18 risk factors were finally identified. A structured survey (Appendix A) was first undertaken. The survey contains all cloud computing risk factors discussed in Section 3.4. The survey was undertaken in order to identify the most important risk factors that can affect cloud computing adoption, and to determine which factors have an important effect in the organization's objectives to include and add it to the identified risk factors. In the survey the participants was requested to categorize risk factors into three levels (Important, Neutral, and Not important) according to their effect on cloud computing environment. 35 international experts from different countries (France, India, Jordan, China, KSA, etc..) responded to the survey and all of them agreed that the previously defined factors are important. This survey provides a general evaluation of risk factors related to cloud computing environment. Depending on the expert opinion, all 18 risk factors are considered as input variables to formulate the dataset with one output, which is considered as a the estimated risk. Next each variable is granulated as low, medium, high, and very high. Next, to assign numeric values to each variable we apply one of the data measurement methods called interval scale; each variable has a numerical range value. Risk Factors and their numerical ranges are illustrated in Table 5.1.

Table 5.1. Risk factors associated with their interval values

Risk Factor	Range value	**Risk Factor**	Range value
DI	0 – 3	*R*	1 – 3
IDD	1 – 3	*RE*	0 – 2
RC	0 – 1	*SLA*	0 – 3
BC& SA	1 – 3	*A& AC*	0 – 3
TPM	0 – 2	*ShE*	1 – 3
I& P	0 – 1	*DB*	0 – 2
DL	0 – 3	*DS*	0 – 1
IAP	0 – 1	*VV*	1 – 3
DL& IS	0 – 3	*DI*	0 - 2

Then, 40 expert rules were formulated to link all the 18 input variables and the output. Below is the first rule as an example (the whole set of rules is illustrated in Appendix B):

If RF1 = 0 and RF2 is 1 and RF3 = 0 and RF4 = 1 and RF5 = 0 and RF6 = 0 and RF7 = 0 and RF8 = 0 and RF9 = 0 and RF10 = 1 and RF11 = 0 and RF12 = 0 and RF13 = 0 and RF14 = 1 and RF15 = 0 and RF16 = 0 and RF17 = 1 and RF18 = 0 then risk = 0

Then we use simple linear interpolation to generate data between the 40 rules. We generated 50 data samples (between each rule) using appropriate step sizes (as the assigned values for different variables were different). The collected dataset contains 18 input attributes, which represent the identified risk factors comprising 1951 instances and one output which represent the risk value. Input attributes were labeled as data transfer (DT), insufficient due diligence (IDD), regulatory compliance (RC), business continuity and service availability (BCSA), third party management (TPM), interoperability and portability (IP), data loss (DL), insecure application programming (IAP), data location and Investigative Support (DLIS), recovery (RY), resource exhaustion (RE), service level agreement (SLA), authentication and access control (AAC), shared environment (ShE), data breaches (DB), data segregation (DS), virtualization vulnerabilities (VV) and data integrity (DI).

5.2 Implement Feature Selection Methods

After preparing our dataset, we need to reduce dimensionality of the data, which enables the data mining algorithm to operate faster and more effectively. In this work, feature selection methods were used to accomplish this task and the new selected data sets are shown in Table 5.2. This work is carried out with the help of WEKA software, which provides an implementation for feature subset selection methods. More details about WEKA tool can be found in [160, 161]. The feature selection methods used were: best-first, random search and ranker. These methods are explained herein below:

Best-first search: The best-first search starts with an empty set of features and generates all possible single feature expansions. Then, the subset with the highest evaluation is chosen and is expanded in the same manner by adding single features. If expanding a subset results in no improvement, the best first search can back track to the more promising previous subset and continuous from there. Given enough time, the best first search will explore the entire search space, so it is common to use stopping criterion [132].

Random search: the random search algorithm [162], first randomly select subset, then continues in two different ways. One of them is to follow sequential search. The second is to continue randomly and generate the next subset randomly.

Ranking: consider a set of n examples (X_k, Y_k) (k = 1, ……., n) consisting of m input variables $X_{k,i}$ (I = 1, ……., m) and one output variable Y_k. Ranking makes use of a scoring function S(i) computed from the values $X_{k,i}$ and Y_k, k = 1, ….., n). By convention, we assume that a high score is indicative of a valuable variable and that we sort variables in decreasing order of S(i). Ranking is a filter method. It is preferable to other feature subset selection methods because of its computational and statistical scalability [163].

Table 5.2 New subset using best first method

The Data	Number of Attributes	Name of Attributes
First dataset	4	IDD, DL, DL&IS

Table 5.3 New subset using random search method

The Data	Number of Attributes	Name of Attributes
Second dataset	5	RC, DL, DL&IS, VV, op
Third dataset	10	IDD, RC, BC&SA, I&P, RE, SAL, A& Ac, DB, DI

Table 5.4 New subset using ranker method

The Data	Number of Attributes	Name of Attributes
Fourth dataset	17	DL&IS, A& Ac, DT, VV, R, BC&SA, TPM, DB, DI, RE, DL, SLA, I&P, DS, RC, ShE
Fifth dataset	15	DL&IS, A& Ac, DT, VV, R, BC&SA, TPM, DB, DI, RE, DL, SLA, I&P, DS
Sixth dataset	13	DL&IS, A& Ac, DT, VV, R, BC&SA, TPM, DB, DI, RE, DL, SLA
Seventh dataset	19	DL&IS, A& Ac, DT, VV, R, BC&SA, TPM, DB, DI, RE, DL, SLA, IAP, IDD

After finishing the preprocessing of the dataset, then the obtained datasets are used to build and test the data mining algorithms. In our search to build a light model two datasets were used that have less number of attributes. Datasets are named first dataset, and second dataset with 3, and 4 attributes respectively. Then a typical split was applied to the available data. The samples distribution in training data and test data for each dataset is illustrated in Table 5.5.

Table 5.5 training and test dataset Percentage split

Split-name	Training	testing
A	60%	40%
B	70%	30%
C	80%	20%
D	90%	10%

5.3 Implement Machine Learning Algorithms

In this section, the course of constructing and applying the models to the two preprocessed datasets are presented. Specifically, as discussed in the previous Chapter,

first the model was built using individual learning algorithms and then an ensemble method is used to perform the regression task. The implementation work is mainly done using WEKA.

5.3.1 Individual Machine Learning Algorithm

Throughout this thesis, six algorithms are used as base algorithms for estimating the risk factors associated with cloud computing environment. These algorithms are Extremely Randomized Decision Trees, Instance-Based Knowledge (IBK), Multilayered Perceptron, K*, Isotonic Regression, and Randomizable Filter Classifier. These algorithms are well known in the data mining community and have proved popular in practice. These algorithms are employed from WEKA software with the default setting.

Extremely Randomized Decision Trees: Decision trees (DT) induction algorithm, was proposed by [164]. The Extremely Randomized Decision Trees or extra tree builds an ensemble of unpruned decision or regression trees according to the classical top-down procedure. Extra tree splits the data totally or partially random. Its two main differences with other decision tree induction algorithms are that it splits nodes by choosing cut-points fully at random and that it uses the whole learning sample to grow the trees [35] [164]. This leads to reduced complexity of the induction process, increased speed of training, and weakened correlation between the induced decision trees [165].

Multilayered Perceptron: A popular feed forward (allow signals to travel one way only, from input to output) neural network architecture that maps sets of input data onto a set of appropriate outputs. It consists of a large number of neurons joined together in pattern of connection These units are usually segregated into three classes: input neurons, which receive information to be processed; output neurons, where the results of the processing are found; the middle neurons, known as hidden neurons which detect features exiting in input data and pass the features to the output neurons [127, 133].

K* (Kstar): An instance based learner algorithm,uses the entropy distance measure to measure the distance between two instance. The entropy distance measure has several features; it provides a consistent approach to handling of symbolic attributes, real values attributes, and missing values [141].

Instance-Based Knowledge (IBK): IBK is an Instance-Based Learning method. IBK in its representation it does not derive a rule set or decision tree and storing it, instead, it uses the instances themselves to represent what they learned. Once a set of training instances has been memorized, on encountering a new instance the memory is searched for the training instance [123]. To compare each unseen instance with existing ones, IBK algorithm use distance metric; most commonly Euclidean distance where the Euclidean distance between two points, X = (x_1, x_2, ……., x_n) and Y = (y_1, y_2, ……., y_n) is:

$$d(x, y) = \sqrt{\sum_{i-1}^{n}(x_i - y_i)^2} \qquad (5.1)$$

and the closest existing instance is used to assign the class for the test sample This is considered as the principle of this algorithm [125].

Isotonic regression: Isotonic regression is a regression method which evaluates linear regression models by the weighed least squares [144]. isotonic regression is a linear regression extension, which can be classified as one of the static regression methods. Linear regression is the most commonly used method for regression analysis, which tries to fit the input-output tuples by linear functions. Usually, simple linear regression refers to the regression of *y* on only one output variable *x*; multiple linear regression refers to the case where *y* depends on many input variables $x_1, x_2, \ldots\ldots, x_n$ [166].

Randomizable Filter Classifier: Typically used for running an arbitrary classifier on data that has been passed through an arbitrary filter. Like the classifier, the structure of the filter is based exclusively on the training data and test instances will be processed by the filter without changing their structure [167].

5.3.2 Ensemble of Machine Learning Algorithm

As we have discussed earlier, an ensemble is a set of learning machines the decisions of which are combined to improve the performance of the overall system. After applying the machine learning algorithms to our two dataset we combine them to form our ensemble model. In our experiments we use vote algorithm to construct the ensemble model.

Vote algorithm: Is a class used to combine multiple predictors. Different combinations of probability estimates for regression are available. In vote method each predictor gets one vote, and the majority wins [168]. The vote algorithm is applied with the use of the average of probability method as a combination rule.

5.3.3 Individual ANFIS Models

In this Section, we present the constructing and applying of the ANFIS model to the preprocessed datasets. We build several individual ANFIS models. The implementation work is mainly done through programming with MATLAB 2014.

For the generation of the Sugeno fuzzy inference systems (FIS), ANFIS model is built using grid partitioning and it tuned to run 100 epoch. Five types of membership functions [169] were used to represent each input: Triangular, trapezoidal, Generalized bell, Gaussian, and Guassian2. These functions are listed below:

Triangular MF (TriMF): The triangular curve is a function of a vector, x, and depends on three scalar parameters *a*, *b*, and *c*, as given by

$$f(x; a, b, c) = \max\left(\min\left(\frac{x-a}{b-a}, \frac{c-x}{c-b}\right), 0\right) \qquad (5.2)$$

The parameters *a* and *c* locate the "feet" of the triangle and the parameter *b* locates the peak.

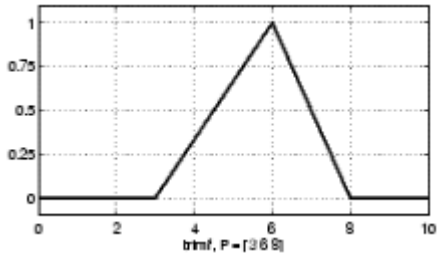

Figure 5.2 Triangular membership function

Trapezoidal MF (TrapMF): The trapezoidal curve is a function of a vector, x, and depends on four scalar parameters *a*, *b*, *c*, and *d*, as given by

$$f(x; a, b, c, d) = \max\left(\min\left(\frac{x-a}{b-a}, 1, \frac{d-x}{d-c}\right), 0\right) \qquad (5.3)$$

The parameters *a* and *d* locate the "feet" of the trapezoid and the parameters *b* and *c* locate the "shoulders."

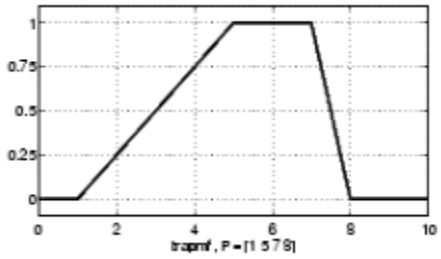

Figure 5.3 Trapezoidal membership function

Generalized bell MF (gbell): The generalized bell function depends on three parameters *a, b,* and *c* as given by

$$f(x; a, b, c) = \frac{1}{1+\left|\frac{x-c}{a}\right|^{2b}} \qquad (5.4)$$

where the parameter *b* is usually positive. The parameter *c* locates the center of the curve. Enter the parameter vector params, the second argument for gbellmf, as the vector whose entries are *a*, *b*, and *c*, respectively.

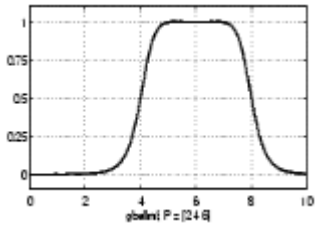

Figure 5.4 Generalized bell membership function

Gaussian MF (gauss): The symmetric Gaussian function depends on two parameters σ and *c* as given by

$$f(x; \sigma, c) = e^{\frac{-(x-c)^2}{2\sigma^2}} \qquad (5.5)$$

The parameters for gaussmf represent the parameters σ and *c* listed in order in the vector [sig c].

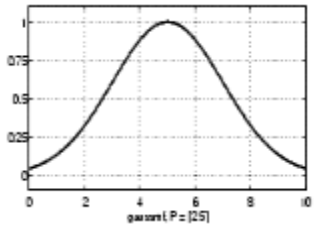

Figure 5.5 Gaussian membership function

Gaussian 2MF (gauss2): A two-sided version of Gaussian membership function. The Gaussian function depends on two parameters *sig* and *c* as illustrated above.

The function gauss2mf is a combination of two of these two parameters. The first function, specified by *sig1* and *c1*, determines the shape of the left-most curve. The second function specified by *sig2* and *c2* determines the shape of the right-most curve. Whenever *c1* < *c2*, the gauss2mf function reaches a maximum value of 1. Otherwise, the maximum value is less than one. The parameters are listed in the order: [sig1, c1, sig2, c2].

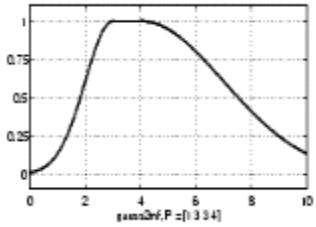

Figure 5.5 A two-sided Gaussian membership function (Gauss2mf)

To determine the structure of our ANFIS model we should decide the fuzzy sets for each input variable. As we mentioned earlier we have two datasets with 3 and 4 input variables. We use two and three fuzzy sets for each input variable. Next, the ANFIS was trained and tested using the training and testing datasets that are discussed in Section 5.2, and the training epoch was set to 100.

5.3.4 Ensemble of ANFIS Models

An individual ANFIS model may or may not offer the best solution. ANFIS employ only a single fuzzy inference system and an ensemble of ANFIS is investigated by the combination of M networks. In-fact, generalization may not be achieved by using a single model. In an ensemble model, a combined decision of many predictors gives us a generalized solution [170].

For a regression problem, ensemble decisions are obtained by averaging the decisions of candidate predictors. However, an average decision lacks in providing the due credit to the best predictors in an ensemble system. Therefore, a weighted average is an alternative, where each predictor in an ensemble system is pre-assigned a weight according to their accuracy/credibility. Now, we have option to assign weight according to the predictor's accuracy. However, in this way, we will lose insights of predictor's decision. Hence, we have used an evolutionary algorithm method for computing the weights of each predictors [171] in the proposed ensemble of ANFIS system. An evolutionary algorithm applies the principles of evolution found in nature to the problem of finding an optimal solution to a Solver problem. In an evolutionary method, we start by initializing random weights to the predictors ranging from -1 to 1. During the evolutionary process, a predictor may acquire negative and positive weight according to its credibility in the ensemble decision. The fitness of the ensemble system having k predictors was computed as:

$$RMSE^{F'}(w_1, w_2, \cdots, w_k) = \sqrt{\frac{1}{N}\sum_i^N \left(\left(\sum_j^k w_j x_{ij}\right) - y_i\right)^2}, \qquad (5.6)$$

Where x_{ij}, is ith decision of jth predictor and y_i denotes target output in learning set that consists of a total of N samples. In the present work, the evolutionary algorithm is used [172] for searching weights $w_1, w_2, \cdots, w_5$ of predictors.

To fulfill the mentioned objective, we need to obtain the best combination of predictor weights. the evolutionary algorithm is used with population size 20, crossover: 0.8, mutation 0.2. The evolutionary based algorithm processes population of possible solutions encoded in form of chromosomes which represent set of weights.

5.4 Summary

This chapter provided the research methodology. It starts by a detailed description of the way followed to simulate the dataset and the feature selection methods used to remove irrelevant features to produce the final datasets used in the experiments. The Chapter then introduced the description and implementation of the machine learning algorithms, and ensemble methods on the preprocessed datasets to build the risk assessment model.

Chapter Six

Experimental Results and Discussions

In this Chapter, the empirical results and the performance of applying the data mining techniques are presented.

6.1 Individual Machine Learning Algorithm Results

The data analysis and the building of the model were carried out using WEKA [160, 161] software environment for machine learning. WEKA is open-source software developed by the University of Waikato and issued under the GNU General Public License. a collection of machine learning algorithms for data mining tasks. Performance statistics are calculated across all datasets using Root Mean Square Error (RMSE) and Correlation Coefficient (CC) but since CC is almost (0.9999 or 1) we did not include them in the tables. We apply attribute selection method to reduce the number of the attributes. In the preprocessing step, the data is filtered to remove the irrelevant data and improve the quality.

Tables from 6.1 through 6.6 show the results of the implementation of data mining algorithms, the best result derived from each algorithm is highlighted. From Table 6.2, we may conclude that multilayer algorithm has the best performance in the case of first dataset with 3 inputs variables. In contrast, the k nearest neighbor in Table 6.4 appears to have the worst performance among all algorithms when applied to the two dataset. Tables 6.1, 6.3, 6.5, and 6.3 show that the isotonic regression, instance based knowledge, random filter classifier and Extremely Randomized Decision Trees algorithms respectively, have the same performance in the both datasets.

Figures 6.1, and 6.2 summarize the performance of each algorithm in the first and second datasets.

Table 6.1 Isotonic regression with first and second datasets

Dataset	A	B	C	D
First dataset	0.0021	**0.0019**	0.002	0.002
Second dataset	0.0021	**0.0019**	0.002	0.002

Table 6.2 Multilayer perceptron with first dataset and second dataset

Dataset	A	B	C	D
First dataset	0.0022	0.0015	0.005	**0.0006**
Second dataset	**0.0019**	0.0024	0.002	0.002

Table 6.3 instance-based knowledge with first dataset and second dataset

Dataset	A	B	C	D
First dataset	0.002	0.0018	0.0018	**0.0017**
Second dataset	0.0021	0.0019	0.0018	**0.0017**

Table 6.4 K* with first dataset and second dataset

	A	B	C	D
First dataset	0.0181	0.019	**0.018**	**0.018**
Second dataset	0.0103	0.011	**0.009**	**0.009**

Table 6.5 Randomizable filter classifier with first dataset and second dataset

Dataset	A	B	C	D
First dataset	0.0021	**0.002**	**0.002**	**0.002**
Second dataset	0.0021	**0.002**	**0.002**	**0.002**

Table 6.6 Extremely Randomized Decision Trees with first dataset and second dataset

Dataset	A	B	C	D
First dataset	0.0043	0.004	**0.003**	**0.003**
Second dataset	0.0042	0.004	0.004	**0.003**

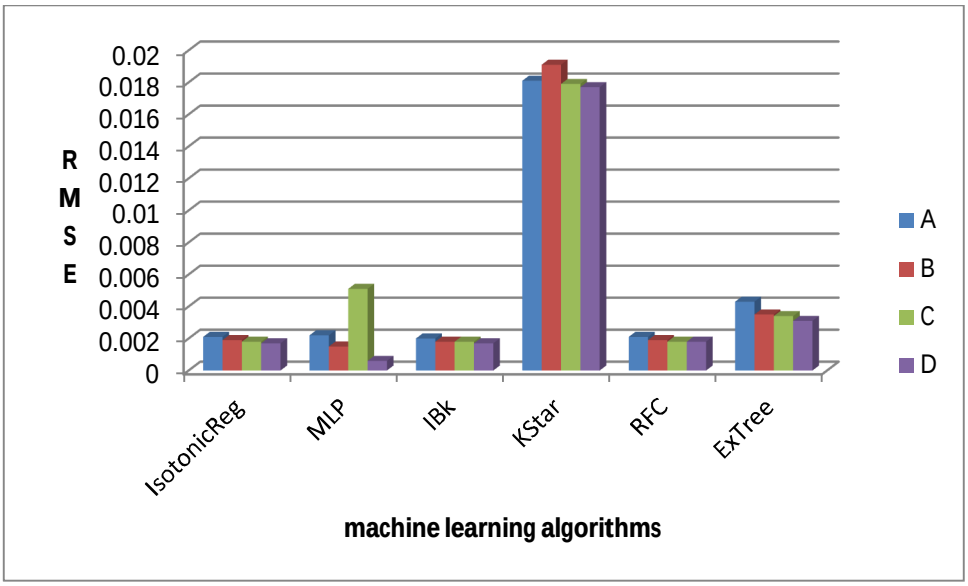

Figure 6.1 Comparison of the machine learning algorithm performance for first dataset

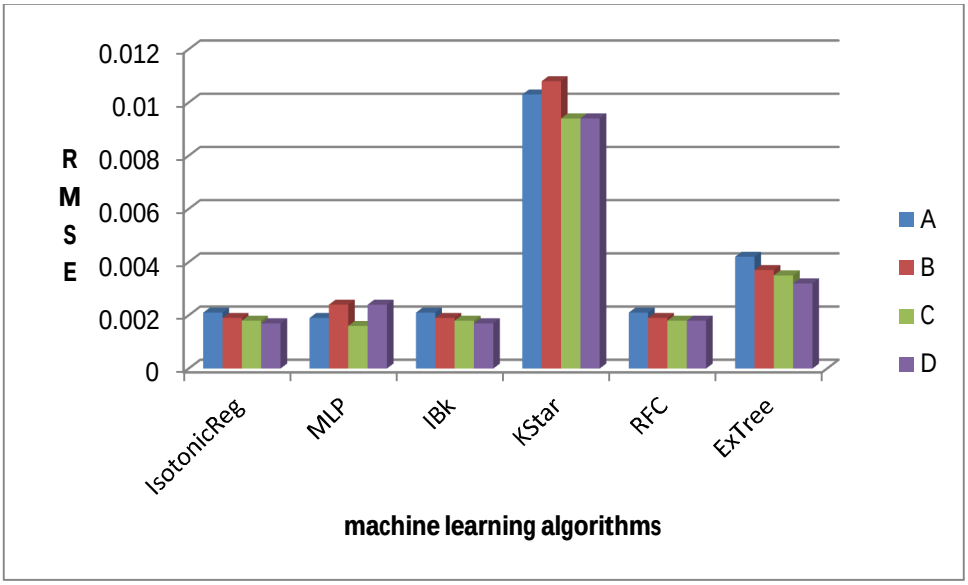

Figure 6.2 Comparison of the machine learning algorithm performance for second dataset

6.2 Ensemble Method Results

Aspiring for better results an ensemble method is used. Ensemble provides for combing the outputs of individual algorithms which leads to improving the performance. Vote combination algorithm is used to combine the algorithms. It uses the majority voting

method as a combination rule as discussed in section 5.3.2. The Tables from 6.7 to 6.9 show results, which were obtained from the ensemble methods. All possible combinations of algorithms were done. The best RMSE value is 0.0009 from first dataset from the combination of 2, 3, and 4 algorithms as it appears in Table 6.10. Figure 6.3 illustrates the results from both datasets with all possible combinations.

Table 6.7 Results of vote algorithm using 2 base algorithms

Algorithms	First dataset	Second dataset
IBK+Isreg	0.0014	0.0013
ET+K*	0.009	0.0048
MLP+RFC	**0.0009**	0.0012
ET+RFC	0.0018	0.0018
IBK+K*	0.0089	0.0048
IBK+ET	0.0019	0.0019
IBK+MLP	0.001	0.0012
IBK+RFC	0.0014	0.0014
Isreg+ET	0.0019	0.0018
Isreg+K*	0.0088	0.0047
Isreg+MLP	**0.0009**	**0.0011**
Isreg+RFC	0.0013	0.0013
ET+MLP	0.0016	0.002
K*+MLP	0.009	0.005
K*+RFC	0.0089	0.0048

Table 6.8 Results of vote algorithm using 3 base algorithms

Algorithm	First dataset	Second dataset
IBK+ K*+ RFC	0.006	0.0033
IBK+ K*+ MLP	0.006	0.0034
IBK+ K*+ ET	0.0061	0.0033

IBK+ K*+ Isreg	0.0014	0.0032
IBK+ RFC+MLP	0.001	0.0011
IBK+RFC+ ET	0.0015	0.0014
IBK+ RFC+Isreg	0.0012	0.0012
IBK+ MLP+ ET	0.0015	0.0015
IBK+ MLP+ Isreg	**0.0009**	**0.001**
IBK+ ET+ Isreg	0.0015	0.0014
K*+ RFC+MLP	0.006	0.0034
K*+ RFC+ ET	0.006	0.0033
K*+ RFC+ Isreg	0.0059	0.0032
K*+ MLP+ ET	0.0061	0.0035
K*+ MLP+Isreg	0.0059	0.0033
K*+ ET+Isreg	0.006	0.0032
RFC+MLP+ ET	0.0012	0.0014
RFC+ MLP+Isreg	0.0009	**0.001**
RFC+ ET+Isreg	0.0014	0.0014
MLP+ ET+Isreg	0.0013	0.0014

Table 6.9 Results of vote algorithm using 4 base algorithms

Algorithm	First data	Second dataset
IBK+ K*+ RFC+ MLP	0.0046	0.0026
IBK+ K*+ RFC+ ET	0.0033	0.0025
IBK+ K*+ RFC+ Isreg	0.0045	0.0025
IBK+ K*+ MLP+ ET	0.0046	0.0026
IBK+K*+ MLP+ Isreg	0.0045	0.0025
IBK+K*+ ET+ Isreg	0.0046	0.0025
IBK+ RFC+ MLP+ET	0.0012	0.0012
IBK+RFC+MLP+Isreg	**0.0009**	**0.001**
IBK+ RFC+ ET+Isreg	0.0013	0.0012

IBK+ MLP+ ET+Isreg	0.0012	0.0012
K*+ RFC+ MLP+ ET	0.0046	0.0026
K*+RFC+MLP+Isreg	0.0045	0.0025
K*+ RFC+ ET+Isreg	0.0045	0.0025
K*+ MLP+ET+Isreg	0.0045	0.0026
RFC+MLP+ET+Isreg	0.0011	0.0012

Table 6.10 Results of vote algorithm using 5 base algorithms

Algorithm	First dataset	Second dataset
IBK+K*+RFC+MLP+ET	0.0037	0.0022
IBK+K*+RFC+MLP+Isreg	0.0036	0.0021
IBK+K*+RFC+ET+Isreg	0.0037	0.002
IBK+Kstar+MLP+ET+Isreg	0.0037	0.0021
IBK+RFC+MLP+ET+Isreg	**0.0011**	**0.0011**
K*+RFC+MLP+ET+Isreg	0.0037	0.0021

Table 6.11 The best results of Vote algorithm

Algorithm	First dataset	Second dataset
2 algorithms	**0.0009**	0.0011
3 algorithms	**0.0009**	**0.001**
4 algorithms	**0.0009**	**0.001**
5 algorithms	0.0011	0.0011

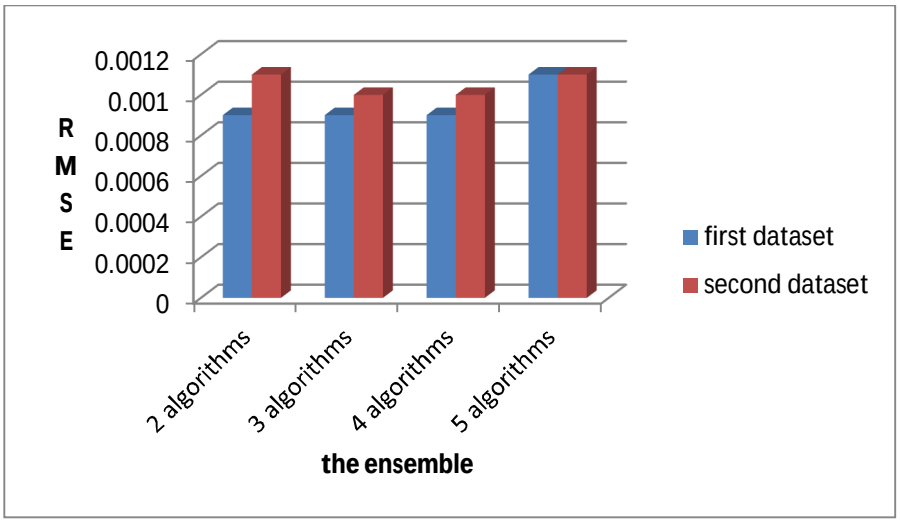

Figure 6.3 The best results of vote algorithm

6.3 Individual ANFIS Models Results

To estimate the risk level, five membership functions (MF) were evaluated. MATLAB ANFIS editor offers different types of MFs including: triangular, trapezoidal, generalized bell (Gbell), Gaussian, and Gaussian 2 which were used in the experiments with 2, and 3 membership functions for each input. Correspondingly all these MFs were evaluated and eventually triangular MF yield the best results with the A subset from both preprocessed dataset, as illustrated in Tables 6.12, and 6.13. All the best results from different MFs are highlighted.

Table 6.12 Individual ANFIS models with 2 fuzzy sets (first dataset)

MF type	**RMSE**			
	A	**B**	**C**	**D**
Tri	**2.882e-06**	3.285e-06	4.022e-06	5.870e-06
Trap	**1.112e-05**	1.320e-05	1.937e-05	3.636e-05
Gbell	**1.874e-05**	2.294e-05	2.787e-05	3.669e-05
Gauss	**1.296e-05**	1.628e-05	2.440e-05	3.466e-05
Gauss2	**1.748e-05**	2.115e-05	2.739e-05	3.868e-05

Table 6.13 Individual ANFIS models with 3 fuzzy sets (first dataset)

MF type	**RMSE**			
	A	**B**	**C**	**D**
Tri	**4.118e-06**	4.886e-06	5.789e-06	7.961e-06
Trap	**1.985e-05**	2.577e-05	3.320e-05	3.320e-05
Gbell	**1.871e-05**	2.115e-05	2.882e-05	3.991e-05
Gauss	**1.976e-05**	2.108e-05	2.840e-05	3.666e-05
Gauss2	**1.977e-05**	2.153e-05	2.826e-05	3.145e-05

Table 6.14 Individual ANFIS models with 2 fuzzy sets (second dataset)

MF type	RMSE			
	A	**B**	**C**	**D**
Tri	**9.150e-06**	1.061e-05	1.407e-05	1.962e-05
Trap	**1.854e-05**	2.262e-05	3.092e-05	4.888e-05
Gbell	**1.801e-05**	2.032e-05	2.397e-05	3.209e-05
Gauss	**1.507e-05**	1.691e-05	2.328e-05	3.201e-05
Gauss2	**1.292e-05**	1.444e-05	1.636e-05	2.505e-05

Table 6.15 Individual ANFIS models with 3 fuzzy sets (second dataset)

MF type	RMSE			
	A	**B**	**C**	**D**
Tri	**9.957e-06**	1.134e-05	1.537e-05	2.163e-05
Trap	**1.766e-05**	2.491e-05	3.454e-05	4.863e-05
Gbell	**1.819e-05**	1.94e-05	2.306e-05	3.392e-05
Gauss	**2.268e-05**	2.727e-05	3.101e-05	4.021e-05
Gauss2	**2.085e-05**	2.381e-05	2.899e-05	5.143e-05

6.4 Ensemble of ANFIS Results

As mentioned in Section 5.4.2 we used evolutionary algorithm to construct the ANFIS ensemble. From the results given in Tables 6.16 to 6.19, it is evident that the ensemble of ANFIS gives the best results with first dataset using subset A when using triangular MF with 2, and 3 membership functions. For the second dataset also with subset A when using triangular MF with 2 and 3 membership functions.

Table 6.16 ANFIS Ensemble for the first dataset with 2 fuzzy sets

Dataset	Predictors Weights					Weighed Ensemble (RMSE)
	Tri	*Trap*	*Gbell*	*Gauss*	*Gauss2*	
A	0.760	0.082	0.057	-0.021	0.122	**1.92408E-06**
B	0.485	0.030	0.240	0.178	0.067	1.21276E-05
C	0.331	0.3175	0.071	0.258	0.023	6.69358E-06
D	0.626	0.187	0.049	0.162	-0.023	9.43936E-06

Table 6.17 ANFIS ensemble for the first dataset with 3 fuzzy sets

Dataset	Predictors Weights					Weighed Ensemble (RMSE)
	Tri	*Trap*	*Gbell*	*Gauss*	*Gauss2*	
A	0.383	0.217	0.114	0.070	0.216	**5.21580E-06**
B	0.519	0.011	-0.008	0.352	0.126	1.09499E-05
C	0.788	0.239	0.248	-0.153	-0.122	8.64922E-06
D	0.534	0.058	0.018	0.114	0.276	1.27589E-05

Table 6.18 ANFIS ensemble for the second dataset with 2 fuzzy sets

Data set	Predictors Weights					Weighed Ensemble (RMSE)
	Tri	Trap	Gbell	Gauss	Guass2	

A	0.702	0.023	0.201	0.032	0.042	**3.21517E-06**
B	0.437	0.334	0.081	0.109	0.039	4.49836E-06
C	0.53	0.05	0.148	0.226	0.046	9.71965E-06
D	0.534	0.058	0.018	0.114	0.276	1.34627E-05

Table 6.19 ANFIS ensemble for the second dataset with 3 fuzzy sets

Data set	Predictors Weights					Weighed Ensemble (RMSE)
	Tri	Trap	Gbell	Gauss	Guass2	
A	0.383	0.217	0.114	0.07	0.216	**5.17103E-06**
B	0.604	0.017	0.112	0.205	0.062	8.30556E-06
C	0.519	0.11	-0.008	0.352	0.126	1.44517E-05
D	0.534	0.058	0.018	0.114	0.276	2.12898E-05

6.5 Discussions

The primary objective of the experiments is to find the lowest RMSE. In these experiments, the advantage of feature selection methods was taken to obtain the best sets of features. The experiments benchmark is RMSE and CC obtained by using first and second datasets. The best results are obtained by Extremely Randomized Decision Trees, Instance-Based Knowledge (IBK), Multilayered Perceptron, K- Nearest Neighbors (K-NN or K*), Isotonic Regression, Randomizable Filter Classifier and vote ensemble algorithm. Performance of ANFIS, and ensemble of ANFIS using evolutionary algorithm are summarized in Table 6.20. The use of the evolutionary algorithm to combine the output of the ANFIS individual models offered the lowest RMSE.

Table 6.20 Best results from all methods

Data mining method	First dataset	Second dataset
	RMSE	
IBK	0.0017	0.0017
KNN	0.018	0.009
RFC	0.002	0.002
MLP	0.0006	0.0019
ET	0.003	0.003
Isreg	0.0019	0.0019
Vote	0.0009	0.001
ANFIS (2 fuzzy sets)	2.882e-06	9.150e-06
ANFIS (3 fuzzy sets)	4.118e-06	9.957e-06
EN-ANFIS (2 fuzzy sets)	1.92408E-06	3.21517E-06
EN-ANFIS (3 fuzzy sets)	5.21580E-06	5.17103E-06

6.6 Summary

This chapter presented all the results of the prediction models built by individual learning algorithms and ensemble methods used in the experiments. In addition, it provided a comparison between all the prediction models results to validate the model.

Chapter Seven

Conclusions

This research introduces risk, risk management, risk assessment definitions, steps, and the importance of risk assessment. A number of risk assessment methods and frame works are applied in the information systems and cloud computing are described. Next, cloud computing system is introduced; emergence, definition, and architecture of cloud computing system are discussed. Furthermore, risk factors associated with cloud computing resources are determined and identified. Next, a description of data mining techniques used to build the proposed model is described. Then, the implementation of those methods is presented and finally, results are provided.

7.1 Thesis contribution

Researchers have different opinions about risk and the association of its dependent variables. Various soft computing tools provide an excellent framework to model risk assessment. The aim of the work presented in this thesis is to increase the chances of cloud computing adoption and to help building trust in the cloud computing services. The main contributions of this thesis are summarized as follows:

- The principles of generic risk assessment have not previously been applied in a formal and structured manner to the field of cloud computing resources risk assessment. In this regard, the development of a generic risk assessment based model for assessing the cloud computing resources risk is considered as novel.
- The development of a practical risk assessment model using data mining techniques to predict the level of risk associated with cloud computing resources. The model was developed following detailed analysis of cloud computing resources risks. The results illustrate effectiveness of the model.
- Use the Ensemble learning techniques and combine individual data mining algorithms outputs to increase efficiency and achieve high accuracy.

7.2 Recommendations

Recommendations for further research focus on the recognized need to assess the repeatability and reliability of aspects of the risk assessment model and associated methodology and on the availability of suitable standards for generating risk category criteria in respect of both qualitative and quantitative variables. There are many ways to further extend the work presented in this thesis. The most appealing ones are listed below:

- The risk assessment model was developed and evaluated using simulated data. Therefore, it would be beneficial to implement the model on a real cloud computing data in order to evaluate the performance. it needs lots of time, funding and more people to work in the team.
- Another extension to the risk model is to consider the internal components of a risk factor rather than considering the factor as a black-box.
- Another area for future work is to use different data mining algorithms to assess the risk level in cloud computing.
- As mentioned earlier, two datasets were used from a total of seven datasets generated by the use of feature selection. Thus, it would be ideal to use this data to evaluate the model.

References

1. Gilb, T. and S. Finzi, *Principles of software engineering management*. Vol. 11. 1988: Addison-Wesley Reading, MA.
2. Gupta, S. and P. Kumar, *Taxonomy of cloud security.* International journal of computer science, engineering and applications, 2013. **3**(5): p. 47.
3. Paquette, S., P.T. Jaeger, and S.C. Wilson, *Identifying the security risks associated with governmental use of cloud computing.* Government Information Quarterly, 2010. **27**(3): p. 245-253.
4. Zhang, Q., L. Cheng, and R. Boutaba, *Cloud computing: state-of-the-art and research challenges.* Journal of internet services and applications, 2010. **1**(1): p. 7-18.
5. Avram, M.-G., *Advantages and challenges of adopting cloud computing from an enterprise perspective.* Procedia Technology, 2014. **12**: p. 529-534.
6. Buyya, R., R. Ranjan, and R.N. Calheiros, *Intercloud: Utility-oriented federation of cloud computing environments for scaling of application services*, in *Algorithms and architectures for parallel processing*. 2010, Springer. p. 13-31.
7. Buyya, R., et al., *Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility.* Future Generation computer systems, 2009. **25**(6): p. 599-616.
8. Subashini, S. and V. Kavitha, *A survey on security issues in service delivery models of cloud computing.* Journal of network and computer applications, 2011. **34**(1): p. 1-11.
9. Ryan, P. and S. Falvey, *Trust in the clouds.* Computer law & security review, 2012. **28**(5): p. 513-521.
10. Zissis, D. and D. Lekkas, *Addressing cloud computing security issues.* Future Generation computer systems, 2012. **28**(3): p. 583-592.
11. Khanmohammadi, K. and S.H. Houmb. *Business process-based information security risk assessment*. in *Network and System Security (NSS), 2010 4th International Conference on*. 2010. IEEE.
12. Shamala, P., R. Ahmad, and M. Yusoff, *A conceptual framework of info structure for information security risk assessment (ISRA).* Journal of Information Security and Applications, 2013. **18**(1): p. 45-52.
13. Saleh, M.S. and A. Alfantookh, *A new comprehensive framework for enterprise information security risk management.* Applied computing and informatics, 2011. **9**(2): p. 107-118.
14. Carroll, M., A. Van Der Merwe, and P. Kotze. *Secure cloud computing: Benefits, risks and controls*. in *Information Security South Africa (ISSA), 2011*. 2011. IEEE.
15. Hu, F., et al., *A review on cloud computing: Design challenges in architecture and security.* CIT. Journal of Computing and Information Technology, 2011. **19**(1): p. 25-55.
16. Parkhill, D.F., *Challenge of the computer utility.* 1966.
17. Kleinrock, L., *A vision for the Internet.* ST Journal of Research, 2005. **2**(1): p. 4-5.

18. Vaquero, L.M., et al., *A break in the clouds: towards a cloud definition.* ACM SIGCOMM Computer Communication Review, 2008. **39**(1): p. 50-55.
19. Buyya, R., C.S. Yeo, and S. Venugopal. *Market-oriented cloud computing: Vision, hype, and reality for delivering it services as computing utilities*. in *High Performance Computing and Communications, 2008. HPCC'08. 10th IEEE International Conference on*. 2008. Ieee.
20. Mell, P. and T. Grance, *The NIST definition of cloud computing.* 2011.
21. Dahbur, K., B. Mohammad, and A.B. Tarakji. *A survey of risks, threats and vulnerabilities in cloud computing*. in *Proceedings of the 2011 International conference on intelligent semantic Web-services and applications*. 2011. ACM.
22. Rimal, B.P., E. Choi, and I. Lumb. *A taxonomy and survey of cloud computing systems.* in *2009 Fifth International Joint Conference on INC, IMS and IDC*. 2009. Ieee.
23. *Kernal Based Virtual Machine*. Available from: www.linux-kvm.org/page/mainpage.
24. *XensourceInc,Xen*. Available from: www.xensource.com.
25. *VMWare ESX server*. Available from: www.vmware.com/product/esx.
26. Bamiah, M.A. and S.N. Brohi, *Seven deadly threats and vulnerabilities in cloud computing.* Int. J. Adv. Eng. Sci. & Techs, 2011(9): p. 87-90.
27. Heiser, J. and M. Nicolett, *Assessing the security risks of cloud computing.* Gartner Report, 2008.
28. Brender, N. and I. Markov, *Risk perception and risk management in cloud computing: Results from a case study of Swiss companies.* International journal of information management, 2013. **33**(5): p. 726-733.
29. Potey, M.M., C. Dhote, and D.H. Sharma, *Cloud Computing-Understanding Risk, Threats, Vulnerability and Controls: A Survey.* International Journal of Computer Applications, 2013. **67**(3).
30. Srinivasamurthy, S. and D.Q. Liu. *Survey on cloud computing security*. in *Proc. Conf. on Cloud Computing, CloudCom*. 2010.
31. Khorshed, M.T., A.S. Ali, and S.A. Wasimi, *A survey on gaps, threat remediation challenges and some thoughts for proactive attack detection in cloud computing.* Future Generation computer systems, 2012. **28**(6): p. 833-851.
32. Samson, T., *9 top threats to cloud computing security.* InfoWorld, 2013.
33. Bhadauria, R., et al., *A survey on security issues in cloud computing.* IEEE Communications Surveys and Tutorials, 2011: p. 1-15.
34. Group, T.T.W., *The notorious nine: cloud computing top threats in 2013.* Cloud Security Alliance, 2013.
35. Brodkin, J., *Gartner: Seven cloud-computing security risks.* Infoworld, 2008. **2008**: p. 1-3.
36. Network, E. and I.S. Agency, *Cloud Computing: Benefits, risks and recommendations for information Security*. 2009: ENISA.
37. Choo, K.-K.R., *Cloud computing: challenges and future directions.* Trends and Issues in Crime and Criminal justice, 2010(400): p. 1.
38. Qian, L., et al., *Cloud computing: an overview*, in *Cloud computing*. 2009, Springer. p. 626-631.

39. Dorey, P. and A. Leite, *Commentary: Cloud computing–A security problem or solution?* information security technical report, 2011. **16**(3): p. 89-96.
40. Armbrust, M., et al., *A view of cloud computing.* Communications of the ACM, 2010. **53**(4): p. 50-58.
41. Luna, J., et al., *Quantitative assessment of cloud security level agreements: A case study.* Proc. of Security and Cryptography, 2012.
42. Bernsmed, K., et al. *Security SLAs for federated cloud services*. in *Availability, Reliability and Security (ARES), 2011 Sixth International Conference on*. 2011. IEEE.
43. Group, U.C.D., *Moving to the Cloud.* 28 February 2011.
44. Raz, T. and D. Hillson, *A comparative review of risk management standards.* Risk Management, 2005: p. 53-66.
45. Van Scoy, R.L., *Software development risk: opportunity, not problem*. 1992, DTIC Document.
46. Roe, P., *A Risk Assessment Based Model for Assessing the Environmental Sustainability of Tourism and Recreation Areas.* 2010.
47. ISO, I., *31000: 2009 Risk management–Principles and guidelines.* International Organization for Standardization, Geneva, Switzerland, 2009.
48. Guide, I., *73: 2009.* Risk management—Vocabulary, 2009.
49. Crouhy, M., D. Galai, and R. Mark, *The essentials of risk management.* Vol. 1. 2006: McGraw-Hill New York.
50. Macdonald, D., *Practical machinery safety*. 2004: Newnes.
51. Djemame, K., et al. *A risk assessment framework and software toolkit for cloud service ecosystems*. in *Proc. 2nd Int. Conf. on Cloud Computing, Grids, and Virtualization*. 2011.
52. López, D., O. Villalba, and L.J. García. *Dynamic risk assessment in information systems: state-of-the-art*. in *Proceedings of the 6th International Conference on Information Technology, Amman*. 2013.
53. Alsoghayer, R.A., *Risk assessment models for resource failure in grid computing*. 2011: University of Leeds.
54. Straub, D.W. and R.J. Welke, *Coping with systems risk: security planning models for management decision making.* Mis Quarterly, 1998: p. 441-469.
55. Stoneburner, G., A. Goguen, and A. Feringa, *Risk Management Guide for Information Technology Systems. 2002.* National Institute of Standards and Technology (NIST), Technology Administration, US Department of Commerce: Gaithersburg, MD, 2009.
56. Fitó, J.O. and J. Guitart, *Business-driven management of infrastructure-level risks in Cloud providers.* Future Generation computer systems, 2014. **32**: p. 41-53.
57. Office, U.G.A., *Information Security Risk Assessment: Practices of Leading Organizations.* 1999.
58. Gjerdrum, D. and M. Peter, *The new international standard on the practice of risk management–A comparison of ISO 31000: 2009 and the COSO ERM framework.* Risk management, 2011(31): p. 8-13.
59. Bartlett, J., *Project risk analysis and management guide*. 2004: APM Publishing Limited.

60. Merna, T. and F.F. Al-Thani, *Corporate risk management*. 2011: John Wiley & Sons.
61. Alberts, C., et al., *Introduction to the OCTAVE Approach.* Pittsburgh, PA, Carnegie Mellon University, 2003.
62. McNally, J.S., *The 2013 COSO Framework & SOX Compliance: One approach to an effective transition.* Strategic Finance, 2013.
63. AIRMIC, A. and A. IRM, *structured approach to Enterprise Risk Management (ERM) and the requirements of ISO 31000.* The Public Risk Management Association, London, UK, 2010.
64. Commission, C.o.S.O.o.t.T., *Internal Control, Integrated Framework.* 2013.
65. Avanesov, E. *Risk management in ISO 9000 series standards*. in *International Conference on Risk Assessment and Management*. 2009.
66. Chaouk, S., *ISO 9001:2015 Information on the revision and insights into the new structure (www.saiglobal.com)*. June 2014.
67. Majstorovic, V.D., *Future Developments of QMS.* Manager (University of Bucharest, Faculty of Business & Administration), 2009(10).
68. ISO, E., *9004: 2009-Managing for the sustained success of an organization--A quality management approach.* International Organization for Standardization, 2009.
69. Alisic, B., *ISO 9004: 2009. A guide towards long term success*. 2008, Recuperado el.
70. Von Solms, B., *Information Security governance: COBIT or ISO 17799 or both?* Computers & Security, 2005. **24**(2): p. 99-104.
71. Ridley, G., J. Young, and P. Carroll. *COBIT and its Utilization: A framework from the literature*. in *System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on*. 2004. IEEE.
72. Al Omari, L., P.H. Barnes, and G. Pitman. *Optimising COBIT 5 for IT governance: examples from the public sector*. in *Proceedings of the ATISR 2012: 2nd International Conference on Applied and Theoretical Information Systems Research (2nd. ATISR2012)*. 2012. Academy of Taiwan Information Systems Research.
73. M.Garsoux, *Cobit 5 ISACA new framework for IT Governance, Risk, Security, and Auditing An overview (http://www.qualified-audit-partners.be/)*. 2013.
74. Preittigun, A., W. Chantatub, and S. Vatanasakdakul, *A Comparison between IT governance research and concepts in COBIT 5.* International Journal of Research in Management & Technology, 2012. **2**(6): p. 581-590.
75. Oliver, D. and J. Lainhart, *COBIT 5: Adding value through effective GEIT.* EDPACS, 2012. **46**(3): p. 1-12.
76. G. Stone, P.N., *Microsoft "Cloud Risk Decision Framework", http://delimiter.com.au/wp-content/uploads/2013/03/SMIC1545_PDF_v7.pdf, accessed on 30/1/2016*.
77. Wickboldt, J.A., et al., *A framework for risk assessment based on analysis of historical information of workflow execution in IT systems.* Computer Networks, 2011. **55**(13): p. 2954-2975.

78. De Bakker, K., A. Boonstra, and H. Wortmann, *Does risk management contribute to IT project success? A meta-analysis of empirical evidence.* International Journal of Project Management, 2010. **28**(5): p. 493-503.
79. van Wyk, R., P. Bowen, and A. Akintoye, *Project risk management practice: The case of a South African utility company.* International journal of project management, 2008. **26**(2): p. 149-163.
80. Kutsch, E. and M. Hall, *Deliberate ignorance in project risk management.* International journal of project management, 2010. **28**(3): p. 245-255.
81. Mkpong-Ruffin, I., et al. *Quantitative software security risk assessment model*. in *Proceedings of the 2007 ACM workshop on Quality of protection*. 2007. ACM.
82. Fredriksen, R., et al., *The CORAS framework for a model-based risk management process*, in *Computer Safety, Reliability and Security*. 2002, Springer. p. 94-105.
83. *http://cloud-standards.org/wiki/index.php?title=Main_Page, accessed on 30-1-2016.*
84. *Cloud standard Customer Council (CSCC), http://cloud-standards.org/wiki/index.php?title=Main_Page, accessed on 30/1/2016*.
85. Stoneburner, G., A. Goguen, and A. Feringa, *Risk Management Guide for Information Technology Systems: Recommendations of the National Institute of Standards and Technology, retrieved November 25, 2009*. 2002.
86. Brunette, G. and R. Mogull, *Security guidance for critical areas of focus in cloud computing v2. 1.* Cloud Security Alliance, 2009: p. 1-76.
87. *CSA Security , trust & Assurance Registry (STAR) https://cloudsecurityalliance.org accessed Jan 16, 2016*.
88. *Cloud Control Matrix (CCM) https://cloudsecurityalliance.org accessed Jan 16, 2016*.
89. *Consensus Assessment Initiative Questionnaire (CAIQ), https://cloudsecurityalliance.org accessed Jan 16, 2016*.
90. Ferrer, A.J., et al., *OPTIMIS: A holistic approach to cloud service provisioning.* Future Generation Computer Systems, 2012. **28**(1): p. 66-77.
91. Djemame, K., et al., *A Risk Assessment Framework for Cloud Computing.* 2014.
92. Saripalli, P. and B. Walters. *Quirc: A quantitative impact and risk assessment framework for cloud security*. in *Cloud Computing (CLOUD), 2010 IEEE 3rd International Conference on*. 2010. Ieee.
93. *NIST, "Standards for Security Categorization of Federal Information and Information Systems FIPS PUB 199" http://csrc.nist.gov/publications/fips/fips199/FIPS-PUB-199-final.pdf accessed on jan 15,2016.* February 2004.
94. Chen, Y., V. Paxson, and R.H. Katz, *What's new about cloud computing security.* University of California, Berkeley Report No. UCB/EECS-2010-5 January, 2010. **20**(2010): p. 2010-5.
95. Linstone, H.A. and M. Turoff, *The Delphi method: Techniques and applications*. Vol. 29. 1975: Addison-Wesley Reading, MA.
96. Fitó, J.O., M. Macías Lloret, and J. Guitart Fernández, *Toward business-driven risk management for cloud computing.* 2010.
97. A. Morali and R. J. wietinga, *Risk -Based Confidentiality Requirements Specification for Outsourced IT Systems (ex-tended version)*. 2010, University of

Twente, 2010: Technical Report TR-CTIT-10-09, Center for Telematics and Information Technology,.
98. Moralı, A. and R. Wieringa. *Risk-based confidentiality requirements specification for outsourced it systems*. in *Requirements Engineering Conference (RE), 2010 18th IEEE International*. 2010. IEEE.
99. Zhang, X., et al. *Information security risk management framework for the cloud computing environments*. in *Computer and Information Technology (CIT), 2010 IEEE 10th International Conference on*. 2010. IEEE.
100. Humphreys, E., *Implementing the ISO/IEC 27001 information security management system standard*. 2007: Artech House, Inc.
101. Stoneburner, G., A.Y. Goguen, and A. Feringa, *Sp 800-30. risk management guide for information technology systems.* 2002.
102. Miller, J., L. Candler, and H. Wald, *Information Security Governance: Government Considerations for the Cloud Computing Environment.* Booz Allen Hamilton, 2009.
103. Cayirci, E., et al. *A Cloud Adoption Risk Assessment Model*. in *Proceedings of the 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing*. 2014. IEEE Computer Society.
104. Peiyu, L. and L. Dong, *The new risk assessment model for information system in cloud computing environment.* Procedia Engineering, 2011. **15**: p. 3200-3204.
105. Kaliski Jr, B.S. and W. Pauley. *Toward risk assessment as a service in cloud environments*. in *Proceedings of the 2nd USENIX conference on Hot topics in cloud computing*. 2010. USENIX Association.
106. Hale, M.L. and R. Gamble. *Secagreement: Advancing security risk calculations in cloud services.* in *Services (SERVICES), 2012 IEEE Eighth World Congress on*. 2012. IEEE.
107. Andrieux, A., et al. *Web services agreement specification (WS-Agreement)*. in *Open Grid Forum*. 2007.
108. Luna, J., et al. *A security metrics framework for the cloud*. in *Security and Cryptography (SECRYPT), 2011 Proceedings of the International Conference on*. 2011. IEEE.
109. Lenkala, S.R., S. Shetty, and K. Xiong. *Security risk assessment of cloud carrier*. in *Cluster, Cloud and Grid Computing (CCGrid), 2013 13th IEEE/ACM International Symposium on*. 2013. IEEE.
110. Wang, H., F. Liu, and H. Liu, *A method of the cloud computing security management risk assessment*, in *Advances in Computer Science and Engineering*. 2012, Springer. p. 609-618.
111. *ISACA Journal, Cloud Computing risk Assessment A case Study, http://www.isaca.org/Journal/archives/2011/Volume-4/Documents/jpdf11v4-Cloud-Computing.pdf, accessed on 30/1/2016.*
112. XIE, X.-m. and Y.-x. ZHAO, *Analysis on the risk of personal cloud computing based on the cloud industry chain.* The Journal of China Universities of Posts and Telecommunications, 2013. **20**: p. 105-112.
113. Almorsy, M., J. Grundy, and A.S. Ibrahim. *Collaboration-based cloud computing security management framework*. in *Cloud Computing (CLOUD), 2011 IEEE International Conference on*. 2011. IEEE.

114. Menzel, M. and C. Meinel. *Securesoa modelling security requirements for service-oriented architectures*. in *Services Computing (SCC), 2010 IEEE International Conference on*. 2010. IEEE.
115. Menzel, M., et al. *The service security lab: A model-driven platform to compose and explore service security in the cloud*. in *Services (SERVICES-1), 2010 6th World Congress on*. 2010. IEEE.
116. Bertram, S., et al. *On-demand dynamic security for risk-based secure collaboration in clouds*. in *Cloud Computing (CLOUD), 2010 IEEE 3rd International Conference on*. 2010. IEEE.
117. Haslum, K., A. Abraham, and S. Knapskog. *Hinfra: Hierarchical neuro-fuzzy learning for online risk assessment*. in *Modeling & Simulation, 2008. AICMS 08. Second Asia International Conference on*. 2008. ieee.
118. Abraham, A., C. Grosan, and V. Snasel. *Programming Risk Assessment Models for Online Security Evaluation Systems*. in *UKSim 2009: 11th International Conference on Computer Modelling and Simulation*. 2009. IEEE.
119. Yucel, G., et al., *A fuzzy risk assessment model for hospital information system implementation.* Expert Systems with Applications, 2012. **39**(1): p. 1211-1218.
120. Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth, *From data mining to knowledge discovery in databases.* AI magazine, 1996. **17**(3): p. 37.
121. Phyu, T.N. *Survey of classification techniques in data mining*. in *Proceedings of the International MultiConference of Engineers and Computer Scientists*. 2009.
122. Kantardzic, M., *Data mining: concepts, models, methods, and algorithms*. 2011: John Wiley & Sons.
123. Chauhan, H., et al. *A comparative study of classification techniques for intrusion detection*. in *Computational and Business Intelligence (ISCBI), 2013 International Symposium on*. 2013. IEEE.
124. Beniwal, S. and J. Arora, *Classification and feature selection techniques in data mining.* International Journal of Engineering Research & Technology (IJERT), 2012. **1**(6).
125. Witten, I.H. and E. Frank, *Data Mining: Practical machine learning tools and techniques*. 2005: Morgan Kaufmann.
126. Edelstein, H.A., *Introduction to data mining and knowledge discovery*. 1998: Two Crows.
127. Guo, R., A. Abraham, and M. Paprzycki, *Analyzing Call Center Performance: A Data Mining Approach.*
128. Pyle, D., *Data preparation for data mining*. Vol. 1. 1999: Morgan Kaufmann.
129. Wu, X., et al., *Top 10 algorithms in data mining.* Knowledge and information systems, 2008. **14**(1): p. 1-37.
130. Liu, H. and L. Yu, *Toward integrating feature selection algorithms for classification and clustering.* Knowledge and Data Engineering, IEEE Transactions on, 2005. **17**(4): p. 491-502.
131. Kittler, J., et al., *On combining classifiers.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, 1998. **20**(3): p. 226-239.
132. Hall, M.A. and L.A. Smith, *Practical feature subset selection for machine learning.* 1998.

133. Kotsiantis, S.B., I. Zaharakis, and P. Pintelas, *Supervised machine learning: A review of classification techniques*. 2007.
134. Hall, M.A., *Correlation-based feature selection for machine learning*. 1999, The University of Waikato.
135. Daelemans, W., et al. *Combined optimization of feature selection and algorithm parameters in machine learning of language*. in *ECML*. 2003. Springer.
136. Karagiannopoulos, M., et al., *Feature selection for regression problems.* Proceedings of the 8th Hellenic European Research on Computer Mathematics & its Applications, Athens, Greece, 2007. **2022**.
137. Raymer, M.L., et al., *Dimensionality reduction using genetic algorithms.* Evolutionary Computation, IEEE Transactions on, 2000. **4**(2): p. 164-171.
138. Saeys, Y., I. Inza, and P. Larrañaga, *A review of feature selection techniques in bioinformatics.* bioinformatics, 2007. **23**(19): p. 2507-2517.
139. Koller, D. and M. Sahami, *Toward optimal feature selection.* 1996.
140. Danso, S.O., *An Exploration of Classification prediction techniques in data mining: the insurance domain.* Master Degree Thesis, Bournemouth University, 2006.
141. Cleary, J.G. and L.E. Trigg. *K*: An instance-based learner using an entropic distance measure*. in *Proceedings of the 12th International Conference on Machine learning*. 1995.
142. Ali, S. and K.A. Smith, *On learning algorithm selection for classification.* Applied Soft Computing, 2006. **6**(2): p. 119-138.
143. Camargo, L. and T. Yoneyama, *Specification of training sets and the number of hidden neurons for multilayer perceptrons.* Neural computation, 2001. **13**(12): p. 2673-2680.
144. Wu, C.-H., W.-H. Su, and Y.-W. Ho, *A study on GPS GDOP approximation using support-vector machines.* Instrumentation and Measurement, IEEE Transactions on, 2011. **60**(1): p. 137-145.
145. Jang, J.-S.R., *ANFIS: adaptive-network-based fuzzy inference system.* Systems, Man and Cybernetics, IEEE Transactions on, 1993. **23**(3): p. 665-685.
146. Buchanan, B.G. and E.H. Shortliffe, *Rule-based expert systems*. Vol. 3. 1984: Addison-Wesley Reading, MA.
147. Wang, Y.-M. and T.M. Elhag, *An adaptive neuro-fuzzy inference system for bridge risk assessment.* Expert Systems with Applications, 2008. **34**(4): p. 3099-3106.
148. Zadeh, L.A., *Fuzzy sets.* Information and control, 1965. **8**(3): p. 338-353.
149. Abraham, A., *Rule-Based Expert Systems.* Handbook of measuring system design, 2005.
150. Takagi, T. and M. Sugeno, *Fuzzy identification of systems and its applications to modeling and control.* Systems, Man and Cybernetics, IEEE Transactions on, 1985(1): p. 116-132.
151. Khoshnevisan, B., et al., *Development of an intelligent system based on ANFIS for predicting wheat grain yield on the basis of energy inputs.* Information processing in agriculture, 2014. **1**(1): p. 14-22.
152. Yang, Z., Y. Liu, and C. Li, *Interpolation of missing wind data based on ANFIS.* Renewable Energy, 2011. **36**(3): p. 993-998.

153. Chang, F.-J. and Y.-T. Chang, *Adaptive neuro-fuzzy inference system for prediction of water level in reservoir.* Advances in Water Resources, 2006. **29**(1): p. 1-10.
154. Negnevitsky, M., *Artificial intelligence: a guide to intelligent systems*. 2005: Pearson Education.
155. Lima, C.A.M.C., ALV Von Zuben, FJ. *Fuzzy systems design via ensembles of ANFIS*. in *Fuzzy Systems, 2002. FUZZ-IEEE'02. Proceedings of the 2002 IEEE International Conference on*. 2002. IEEE.
156. ., R.P., *Ensemble learning.* Scholarpedia,, (2009). **4**: p. 2776.
157. Dietterich, T.G., *Ensemble methods in machine learning*, in *Multiple classifier systems*. 2000, Springer. p. 1-15.
158. Prodromidis, A., P. Chan, and S. Stolfo, *Meta-learning in distributed data mining systems: Issues and approaches.* Advances in distributed and parallel knowledge discovery, 2000. **3**: p. 81-114.
159. Subramanian, S., V.B. Srinivasan, and C. Ramasa, *Study on classification algorithms for network intrusion systems.* Journal of Communication and Computer, 2012. **9**(11): p. 1242-1246.
160. Hall, M., et al., *The WEKA data mining software: an update.* ACM SIGKDD explorations newsletter, 2009. **11**(1): p. 10-18.
161. Witten, I.H., et al., *Weka: Practical machine learning tools and techniques with Java implementations.* 1999.
162. Liu, H.Y., Lei, *Toward integrating feature selection algorithms for classification and clustering.* Knowledge and Data Engineering, IEEE Transactions on, 2005. **17**(4): p. 491-502.
163. Guyon, I. and A. Elisseeff, *An introduction to variable and feature selection.* The Journal of Machine Learning Research, 2003. **3**: p. 1157-1182.
164. Geurts, P., D. Ernst, and L. Wehenkel, *Extremely randomized trees.* Machine learning, 2006. **63**(1): p. 3-42.
165. Désir, C., et al., *Classification of endomicroscopic images of the lung based on random subwindows and extra-trees.* Biomedical Engineering, IEEE Transactions on, 2012. **59**(9): p. 2677-2683.
166. Larose, D.T., *Data mining methods & models*. 2006: John Wiley & Sons.
167. *Randomizable filter classifier*. 20/5/2016]; Available from: http://weka.sourceforge.net/doc.dev/weka/classifiers/meta/package-summary.html. .
168. *Vote*. 20/5/2016]; Available from: http://weka.sourceforge.net/doc.dev/weka/classifiers/meta/package-summary.html.
169. *Adaptive Neuro-Fuzzy Inference System*, . 14/6/2016]; Available from: http://www.mathworks.com/help/fuzzy.
170. Lima, C.A.M., A.L. Coelho, and F.J. Von Zuben. *Fuzzy systems design via ensembles of ANFIS*. in *Fuzzy Systems, 2002. FUZZ-IEEE'02. Proceedings of the 2002 IEEE International Conference on*. 2002. IEEE.
171. Ojha, V.K.J., Konrad Abraham, Ajith Snášel, Václav, *Dimensionality reduction, and function approximation of poly (lactic-co-glycolic acid) micro-and*

nanoparticle dissolution rate. International journal of nanomedicine, 2015. **10**: p. 1119.
172. Goldberg, D.E.H., John H, *Genetic algorithms and machine learning.* Machine learning, 1988. **3**(2): p. 95-99.

Appendix A

Risk Factor Survey

Risk Factors of Cloud Computing

This survey questionnaire is designed to evaluate the security related risk factors in cloud computing environment. We have identified several risk factors as reported in the academic literature, which are considered as the important. This survey is part of a PhD research and the purpose is to evaluate the risk factors, by determining the influence of these factors by categorizing them under three levels:

Important: If the evaluated factor likely happens, it affects the cloud environment
Neutral: If the evaluated factor is likely to happen, it affects the cloud environment moderately.
Not- important: If the evaluated factor likely happens, it affect the cloud environment very little or negligible.

Risk 1. Authentication and access control:

Organization's private and sensitive data must be secure and only authenticated users can access it. When using cloud, the data is processed and stored outside the premise of an enterprise, which brings a level of risk because outsourced services bypass the "physical, logical, and personnel controls", any outside or unwanted access is denied

- ◯ Important
- ◯ Neural
- ◯ Not Important

Risk 2. Data loss:

Data loss means that the valuable data disappear into the ether without a trace, cloud customers need to make sure that this will never happen to their sensitive data.

- ◯ Important
- ◯ Neural
- ◯ Not Important

Risk 3. Insecure Application Programming Interface:

APIs is an important and necessary part to security and availability of whole cloud services. Building interfaces, injecting services will increase risk, there for some organization may in force to relinquish their credentials to third party in order to enable their agency

- ◯ Important
- ◯ Neural
- ◯ Not Important

Risk 4. Network and internet:

sensitive data is obtained from customers, processed and stored at cloud provider end. All data flow over network needs to be secured in order to prevent seepage of customer's sensitive information. The application provided by cloud provider to their customers is has to be used and managed over the web. The risk come from the security holes in the web application.

- ◯ Important
- ◯ Neural
- ◯ Not Important

Risk 5. Insufficient due diligence:

before start using cloud services, the organization need to fully understand the cloud environment and its associated risk.

- ◯ Important
- ◯ Neural
- ◯ Not Important

Risk 6. Shared environment:

Multi-tenancy is key factor of cloud computing service. To achieve scalability cloud provider provide shared infrastructure, platform, and application to deliver their services, this shared nature enable multiple users to share same computer resources, which may lead to leaking data to other tenants, also, if one tenant carried malicious activities the reputation of other tenants may be affected. The impact can be appear as a problems for the organization's reputation in addition to service delivery, and data loss.

- ◯ Important
- ◯ Neural
- ◯ Not Important

Risk 7. Regulatory compliance:

If the provider is unable or unwilling to subjected to external audits and security certification, and they donot give their customers some information about the security controls that have been evaluated. it should only be considered for most trivial functions. Regardless of location, the custodian is ultimately responsible for ensuring the security, protection, and integrity of the data, especially when they are passed to a third party.

- ◯ Important
- ◯ Neural
- ◯ Not Important

Risk 8. Data breaches:

Breaching in to cloud environment will potentially attack all users data. Those attackers can exploit a single flaw in one client's application to get to all other client's data as well, if the cloud service databases are not designed properly.

- ◯ Important
- ◯ Neural
- ◯ Not Important

Risk 9. Business continuity and service availability:

The nature of business environment, competitive pressure, and the changes happening in it leads to some events that may affect the cloud service provider, such as merger, go broke, bankruptcy, or it acquisition by another company. These things lead to loss or deterioration of service delivery performance, and quality of service. Another important thing to the cloud computing provider is that their customers must be provided with service around the clock, but outages do occur and can be unexpected and costly to customers.

- ◯ Important
- ◯ Neural
- ◯ Not Important

Risk 10. Data location and investigative support:

Most cloud service providers have many data centers around the globe. When regards to privacy regulation in different jurisdiction, in different countries where the government restrict the access to data in their borders, or if the data stored in high-risk countries, all these things make data location big concern issue. The investigation of an illegal activity may be impossible in cloud computing environment, because multiple customer's data can be a located in different data centers that are spread around the globe, If the enterprise relies on the cloud service for the processing of business records then it must take into account the factor of inability or unwillingness of the provider to support it.

- ◯ Important
- ◯ Neural
- ◯ Not Important

Risk 11. Data segregation:

The risk arise here come from the failure of the mechanisms to separate data in storage, and memory, from multiple tenants in the shared infrastructure.

- ◯ Important
- ◯ Neural
- ◯ Not Important

Risk 12. Recovery:

Cloud users do not know where their data is hosted. Some events such as man-made, or natural disaster may happen; in such events customers need to know what will their data and long the recovery process take.

- ◯ Important
- ◯ Neural
- ◯ Not Important

Risk 13. Virtualization vulnerabilities:

Virtualization is one of the fundamental components of the cloud service. However it introduces major risks as every cloud provider uses it. Beside its own risks it hold every risk posed by physical machines.

- ◯ Important
- ◯ Neural
- ◯ Not Important

:Risk 14. Third part management:

There are many issues in cloud computing related to third party because the client organizations are not directly managed by the cloud service provider. Some old concerns in information security appear with outsourcing such as integrity control and sustainability of supplier and all risks that client may take if it rely on a third party.

- ◯ Important
- ◯ Neural
- ◯ Not Important

Risk 15. Interoperability and portability:

Interoperability and portability become crucial because if the organization locks to a specific cloud provider, then the organization will be at the mercy of the service level and pricing policies of that provider and it hasn't the freedom to work with multiple cloud provider.

- ◯ Important
- ◯ Neural
- ◯ Not Important

Risk 16. Resource exhaustion:

Cloud provider allocates resource according to statistical projections. Inaccurate modeling of resources usage can lead to many issues such as: service unavailability, access control compromised, economic and reputational losses, and infrastructure oversize.

- ◯ Important
- ◯ Neural
- ◯ Not Important

Risk 17. Service level agreement:

The organization needs to ensure that the terms of (SLA) are being met. Risk may appear with service level application such as the data owner as some cloud provider include explicitly some terms state that the data stored is the provider's not the customer's. In few cases where cloud vendor went out of business, their customer private data sold as part of the asset to the next buyer. Also (SLA) terms should include Licensing conditions, there is the possibility for creating original work in the cloud, but if not protected by the appropriate contractual clauses, this original work may be at risk. One of the (SLA) terms must be for responsibilities of cloud provider for enabling governance.

- ◯ Important
- ◯ Neural
- ◯ Not Important

Risk 18. Data integrity:

One of the most critical elements in all systems is data integrity. Cloud computing magnified the problem of data integrity,1. The biggest challenge, which endanger the data integrity is transaction management, at the protocol level, does not support transactions or guaranteed delivery. If data integrity is not guaranteed and there is lack in integrity controls, this may result in deep problems.

- ◯ Important
- ◯ Neural
- ◯ Not Important

Appendix B

The Rules (R1 – R40)

Risk factor	DT	IDD	RC	BC&SA	TPM	I&P	DL	IAP	DL&IS	R	RE	SLA	A&AC	ShE	DB	DS	VV	DI	RL
R1	0	1	0	1	0	0	0	0	0	1	0	0	1	1	0	0	1	0	0
R2	0.008	1.003	0.001	1.006	0.003	0.002	0.008	0.001	0.006	1.003	0.003	0.008	1.003	1.003	0.002	0.002	1.003	0.003	0.005
R3	0.028	1.006	0.003	1.017	0.006	0.004	0.028	0.003	0.028	1.006	0.006	0.028	1.006	1.005	0.004	0.004	1.005	0.006	0.008
R4	0.047	1.009	0.005	1.026	0.009	0.006	0.047	0.006	0.046	1.009	0.009	0.047	1.009	1.008	0.006	0.006	1.008	0.009	0.025
R5	0.066	1.025	0.007	1.035	0.025	0.008	0.066	0.008	0.066	1.025	0.025	0.066	1.025	1.025	0.008	0.008	1.025	0.025	0.045
R6	0.086	1.043	0.009	1.044	0.043	0.009	0.086	0.009	0.082	1.043	0.043	0.086	1.043	1.043	0.025	0.009	1.043	0.043	0.066
R7	0.17	1.059	0.013	1.057	0.059	0.01	0.17	0.012	0.17	1.058	0.059	0.17	1.059	1.057	0.043	0.01	1.057	0.059	0.086
R8	0.37	1.071	0.019	1.068	0.069	0.019	0.37	0.019	0.37	1.075	0.071	0.37	1.071	1.071	0.059	0.019	1.071	0.071	0.17
R9	0.58	1.082	0.026	1.077	0.079	0.025	0.58	0.026	0.58	1.084	0.079	0.58	1.085	1.085	0.075	0.025	1.084	0.079	0.37
R10	0.74	1.093	0.031	1.085	0.086	0.031	0.74	0.031	0.84	1.097	0.086	0.74	1.095	1.095	0.84	0.031	1.097	0.086	0.58
R11	0.95	1.12	0.038	1.094	0.095	0.038	0.99	0.038	0.99	1.12	0.095	0.99	1.12	1.12	0.96	0.038	1.12	0.095	0.74
R12	1.027	1.31	0.043	1.15	0.18	0.043	1.005	0.043	1.005	1.31	0.12	1.005	1.31	1.25	0.19	0.043	1.31	0.12	0.99
R13	1.045	1.39	0.049	1.26	0.35	0.049	1.027	0.049	1.027	1.39	0.35	1.027	1.39	1.39	0.35	0.049	1.39	0.35	1.005
R14	1.066	1.47	0.056	1.35	0.49	0.055	1.042	0.056	1.042	1.45	0.49	1.042	1.45	1.45	0.46	0.055	1.45	0.49	1.027
R15	1.089	1.56	0.061	1.46	0.61	0.061	1.049	0.061	1.049	1.54	0.61	1.049	1.54	1.54	0.58	0.061	1.54	0.61	1.042
R16	1.22	1.64	0.067	1.54	0.75	0.067	1.058	0.067	1.057	1.61	0.75	1.058	1.61	1.61	0.69	0.067	1.61	0.75	1.049
R17	1.35	1.73	0.074	1.65	0.83	0.074	1.066	0.074	1.064	1.75	0.83	1.066	1.75	1.73	0.83	0.074	1.75	0.83	1.056
R18	1.42	1.81	0.082	1.74	0.91	0.081	1.072	0.082	1.072	1.82	0.91	1.072	1.82	1.81	0.91	0.081	1.82	0.91	1.065
R19	1.55	1.89	0.088	1.84	0.97	0.088	1.097	0.087	1.097	1.89	0.97	1.097	1.89	1.89	0.97	0.088	1.89	0.97	1.072
R20	1.67	1.95	0.093	1.93	1	0.094	1.2	0.093	1.2	1.97	1	1.3	1.97	1.97	1	0.094	1.95	1	1.097

R21	1.74	2.001	0.098	2.005	1.007	0.098	1.4	0.097	1.4	2.001	1.004	1.5	2.001	2.001	1.004	0.098	2.001	1.003	1.3
R22	1.82	2.003	0.11	2.013	1.015	0.1	1.6	0.1	1.6	2.003	1.006	1.7	2.003	2.003	1.006	0.11	2.003	1.006	1.5
R23	1.93	2.006	0.15	2.024	1.024	0.14	1.8	0.15	1.8	2.006	1.009	1.9	2.006	2.005	1.009	0.15	2.006	1.009	1.7
R24	2.001	2.008	0.19	2.035	1.032	0.19	2.001	0.19	2	2.008	1.025	2.001	2.008	2.007	1.025	0.19	2.008	1.025	1.9
R25	2.003	2.019	0.23	2.044	1.043	0.23	2.003	0.23	2.002	2.019	1.043	2.003	2.019	2.019	1.043	0.23	2.019	1.043	2.003
R26	2.005	2.035	0.27	2.056	1.054	0.27	2.005	0.27	2.005	2.035	1.059	2.005	2.035	2.035	1.059	0.27	2.035	1.059	2.005
R27	2.007	2.058	0.31	2.064	1.071	0.31	2.007	0.31	2.007	2.058	1.071	2.007	2.058	2.058	1.071	0.31	2.058	1.071	2.007
R28	2.009	2.073	0.35	2.73	1.085	0.35	2.009	0.35	2.009	2.073	1.085	2.009	2.073	2.073	1.085	0.35	2.073	1.085	2.009
R29	2.014	2.094	0.39	2.084	1.095	0.39	2.015	0.39	2.014	2.094	1.095	2.014	2.094	2.094	1.095	0.39	2.094	1.095	2.015
R30	2.033	2.18	0.45	2.096	1.12	0.44	2.033	0.44	2.033	2.15	1.12	2.033	2.15	2.15	1.12	0.45	2.13	1.12	2.033

R31	2.056	2.33	0.49	2.18	1.31	0.48	2.055	0.48	2.056	2.38	1.31	2.056	2.38	2.38	1.31	0.49	2.37	1.31	2.055
R32	2.074	2.43	0.53	2.27	1.39	0.53	2.074	0.53	2.074	2.44	1.39	2.074	2.44	2.44	1.39	0.53	2.44	1.39	2.074
R33	2.092	2.51	0.59	2.38	1.45	0.58	2.092	0.58	2.092	2.51	1.45	2.092	2.51	2.51	1.45	0.58	2.53	1.45	2.092
R34	2.16	2.59	0.66	2.46	1.54	0.65	2.11	0.65	2.12	2.59	1.54	2.12	2.59	2.59	1.54	0.66	2.59	1.54	2.12
R35	2.29	2.67	0.71	2.55	1.61	0.71	2.28	0.71	2.29	2.66	1.61	2.29	2.66	2.66	1.61	0.71	2.66	1.61	2.28
R36	2.38	2.76	0.76	2.64	1.75	0.76	2.36	0.76	2.38	2.76	1.75	2.38	2.76	2.76	1.75	0.76	2.76	1.75	2.37
R37	2.56	2.83	0.82	2.76	1.86	0.81	2.56	0.81	2.56	2.83	1.86	2.56	2.83	2.83	1.86	0.82	2.82	1.86	2.55
R38	2.73	2.89	0.88	2.84	1.91	0.88	2.75	0.88	2.73	2.89	1.91	2.73	2.89	2.89	1.91	0.88	2.89	1.91	2.75
R39	2.91	2.95	0.96	2.93	1.98	0.95	2.94	0.95	2.94	2.95	1.98	2.94	2.95	2.95	1.98	0.96	2.95	1.98	2.94
R40	3	3	1	3	2	1	3	1	3	3	2	3	3	3	2	1	3	2	3

PUBLICATION

Some parts of the work presented in this thesis have been published in the following articles:

Journal Papers

1- Modeling Security Risk Factors in a Cloud Computing Environment**.** Journal of Network and Innovative Computing ISSN 2160-2174, Volume 1 (2013) pp. 348-355
2- An Ensemble of Neuro-Fuzzy Model for Assessing Risk in Cloud Computing Environment, Nada Ahmed, Varun Kumar Ojha, Ajith Abraham, Journal of Information Assurance and Security, ISSN: 1554-1010, Volume 10, Issue 5, pp. 226-231, 2015

Conference Papers

3- Modeling Cloud Computing Risk Assessment Using Machine Learning. Ahmed, N., & Abraham, A. (2015, January). In *Afro-European Conference for Industrial Advancement* (pp. 315-325). Springer International Publishing.
4- Modeling Cloud Computing Risk Assessment Using Ensemble Methods. Ahmed, N., & Abraham, A. (2015). In *Pattern Analysis, Intelligent Security and the Internet of Things* (pp. 261-274). Springer International Publishing.
5- Neuro-Fuzzy Model for Assessing Risk in Cloud Computing Environment. Ahmed, N., & Abraham, A. In: Afro-European Conference for Industrial Advancement, September 9- 11, 2015, Villejuif, France
6- An Ensemble of Neuro-Fuzzy Model for Assessing Risk in Cloud Computing Environment. Nada Ahmed, Varun Kumar Ojha, Ajith Abraham, in: Nature and Biologically Inspired Computing (NaBIC2015), December 01-03, 2015, Pietermaritzburg, South Africa.