thesis

by Dinara khashimova

Submission date: 08-Jun-2020 08:14PM (UTC+0300)
Submission ID: 1340206526
File name: Thesis_2020_12.pdf (2.41M)
Word count: 13548
Character count: 74543

Ministry of Education and Science of the Republic of Kazakhstan
Suleyman Demirel University

Dinara Khashimova

The prediction of information security level in the enterprise

THESIS

Degree of Master of Science in Mathematics
(degree code: 6M070400)
Department of Mathematics
Faculty of Engineering and Natural Sciences

Supervisor: **Lyazzat Atymtayeva**

Kaskelen, 2020

Abstract

This thesis presents the results of an analysis to identify groups of threats specific to the infrastructure and systems of an enterprise, which is one of the main stages in forecasting. The state of information security at enterprises is considered, the qualifications of security threats and classification methods based on attack methods and the impact of threats are analyzed. Threats for the safe use of the Internet and hacking sites, data theft, phishing attacks and social engineering are assessed; Identification of cloud computing security threats that are encountered in the enterprise's Internet networks. The advantages and disadvantages of Web Application Firewall, which are used to protect attacks, such as DDoS attacks, SQL injections, cross-site scripting, and others, are studied. Works for providing protection using artificial intelligence and machine learning are presented.

Аңдатпа

Бұл зерттеу жұмысында болжаудың негізгі кезеңдерінің бірі болып табылатын кәсіпорынның инфрақұрылымы мен жүйелеріне тән қауіптер тобын анықтау үшін жасалған талдау нәтижелері келтірілген. Кәсіпорындардағы ақпараттық қауіпсіздіктің жай-күйі қарастырылады, қауіп-қатерлердің білік-тілігі, шабуыл әдістеріне және қауіп-қатерлерге негізделген классификация-лау әдістері талданады. Интернетті қауіпсіз пайдалану және сайттарды бұзу, деректерді ұрлау, фишингтік шабуылдар және әлеуметтік инженерия үшін қауіптер бағаланады. Кәсіпорынның Интернет желілерінде кездесетін бұлтты есептеу қауіпсіздігіне қауіптерді анықтау. DDoS шабуылдары, SQL инъекция-лар, сайттар аралық сценарийлер (XSS) және басқалары сияқты шабуылдар-дан қорғау үшін қолданылатын Web Application Firewall артықшылықтары мен кемшіліктері зерттелген. Жасанды интеллект пен машиналық оқытуды қолдана отырып қорғауды қамтамасыз ететін жұмыстар ұсынылған.

Аннотация

В данной работе представлены результаты анализа по выявлению групп угроз, специфичных для инфраструктуры и систем предприятии которое является одним из основных этапов в прогнозировании. Рассмотрены состояние информационной безопасности на предприятиях, проанализированы квалификации угроз безопасности и методов классификации, основанные на методах атак и на воздействии угроз. Оценены угрозы по безопасному использованию Интернета и взламыванию сайтов, краж данных, атакам фишинга и социальной инженерии; выявление угроз безопасности облачных вычислений, которые встречаются в интернет сетях предприятии. Изучены преимущества и недостатки Файрвол веб-приложений, который применяются для защиты атак, такие как DDoS-атаки, SQL-инъекции, межсайтовый скриптинг (XSS), и др. Представлены работы для обеспечения защиты с применением искусственного интеллекта и машинного обучения.

Acknowledgements

At first, I would like to express my gratitude for being my supervisor, associate professor, doctor of physical and mathematical sciences Lyazzat Atymtaeva from the engineering and natural sciences of Suleiman Demirel University. Professor Lyazzat Atymtaeva has always been genious whenever I faced with any troubles or had any requires about the thesis. She supported me in writing this dissertation, moreover guided me all the time.

I would like to praise the experts taking part in doing research project: PhD. Assoc. Professor Olimjon Baimuratov and Assoc. Professor of Engineering and Natural Sciences at the University of Suleyman Demirel Shirali Kadyrov. Without their support and guidance I wouldn't be able to carry out my thesis.

Contents

1. Introduction

1.1 Motivation

Identification of basic machine learning models to detect malicious threats using the firewall of a web application for forecasting information security levels, which will allow enterprises/organizations to reduce costs, accelerate the audit process and improve its quality by bringing it in line with international information security standards.

In the recent couple of years there has been a significant increase of application layer based attacks [31]. According to the Open Web Application Security Project attacks such as Cross site scripting (XSS), SQL injections and DDoS-atacks are the kind of attacks applications are usually most susceptible to [27]. There are several measures to take into account to mitigate these kind of attacks. Sanitizers and Content Security Policies are examples of these protection mechanisms. Another way to mitigate application layer attacks, which this report will focus on, is Web Application Firewalls (WAF) with intrusion detection. With the rapid growth of data, machine learning has been starting to play a significant role within application firewalls.

1.2 Aims and Objectives

The main goal of this work is to investigate the potential of using machine learning methods to detect intrusions at the WAF application level, and also to explore what good and bad functions to use in machine learning classifiers when trying to detect malicious code. In this paper, we try to determine an effective mechanism and methods for analyzing information security based on identifying correctly selected tools and models, such as statistical analysis, visualization techniques,

machine learning methods: the stages of creating a mathematical model.

1.3 Thesis Outline

In modern conditions, it is almost impossible to find such a branch of human activity where information and communication technologies would not be applied. Their rapid development helps to reduce the security of critical information used and increase the demand for tools and ways to protect it. Moreover, the relevance and importance of the problem of ensuring information security is determined by the following factors: - an increase in the volume of information accumulated, stored and processed by computers; - the spread of network technologies and the global Internet; - expanding the circle of users with direct access to computing resources and data arrays.

Taking into account that information security is a multifunctional and very complex system in terms of management, it becomes necessary to use more effective mechanisms to ensure protection and eliminate security threats. The best solution today is to conduct an information security audit in organizations. This process involves analyzing the level of information security of the organization in accordance with various security standards, however, in essence, it is highly costly both in terms of time and cost, and in terms of human resources involvement. Automation of the audit process can reduce costs, accelerate the audit process and improve its quality by bringing it into line with international standards in the field of information security. At present, in the modern world of information technology there is no openly distributed and widely used software for auditing information security systems, since the security audit process is understandably closed to public access. Separate consulting and auditing companies create auxiliary software templates for external and internal audit, which allow reflecting individual elements of system auditing. Existing expert information security audit systems cannot provide reliable information and reflect the real level of security in the organization, since they are based on a fuzzy scale and depend on the size, frequency of use and constant updating of the database and the knowledge base used in the security audit process.

Under such conditions, it becomes necessary to search for a more effective and efficient mechanism for analyzing the level of information security of organizations

that has significant results, despite the rapidly changing world of technology, cyber threats, information, etc. Recently, the direction of DataScience (data science) has become widespread, in which, based only on historical statistical data (the so-called datasets), it is possible to make a forecast and analyze the state of the predicted object by using machine learning algorithms.

1.4 Literature review

When reviewing some research works in the field of analysis and management of organization security, it was found that scientists are interested in methods, tools, effective mechanisms and models for managing information security.

So, researcher R. Leszczyn in his work "Cost assessment of computer security activities" (2013) [25] presents a new method for estimating the cost of computer security measures and tests them with a real example when assessing the cost of computer security measures. The article contains a comprehensive analysis of the costs and benefits of assessing the value of information security, which makes it clear that it is necessary to create a mechanism for assessing the current level of security.

The authors of D. Schatz, R. Bashroush in "Security predictions-A way to reduce uncertainty" (2019) [33] appear a topical modeling approach to recognize 17 key anticipated dangers based on more than 200 security figures distributed in 2015. At the same time, they utilized a study strategy based on quantitative information. Uncovered covered up expectations related programmer political activities, large-scale infringement of individual information and therapeutic records, expanded dangers from different sorts of malware, in specific ran-somware, as well as large-scale DDoS attacks. The article centers on strategies to progress the viability of innovations and procedures of cybersecurity of organizations.

Authors J. Huang. C. Chiang, J. Chang in 2018 published the work "Email security level classification of imbalanced data using artificial neural network: The real case in a world-leading enterprise" [19], in which they considered the issue of guaranteeing mail security. They emphasized that, of course, email is much more helpful than conventional mail in message conveyance, but there's a leak of data that's basic for trade. Agreeing to researchers, this issue can be fathomed by classifying emails at different security levels utilizing innovation for content

mining and machine learning. In this consider, the creators created a plot in which a neural arrange is utilized to extricate data from emails in arrange to guarantee its change into a multidimensional vector. Email is classified using an artificial neural network.

Researchers V. Jaganathan, P. Cherurveettil, P. Sivashanmugam in their work "Using a Prediction Model to Manage Cyber Security Threats" [] (2015) addressed the topic of cyber attacks. They emphasized that cyber-attacks are a major issue facing all organizations. Protecting the organizations' information systems, therefore, plays a crucial role. At the same time, organizations must be able to understand the ecosystem and predict attacks, and the risk management should include quantitative forecasting of attacks. One risk factor is the malware 's effect on the integrity of safe objects. In this article, they propose a mathematical model to predict the effects of an attack based on cyber-security factors. The mathematical model is widespread and can be adapted to an individual organization's needs.

Another "Cyber Attacks Prediction Model Based on Bayesian Network" [39] article on cyber attacks was presented by J. Wu, L. Yin in 2012. They also emphasized that predicting cyberattacks is an important part of risk management. However, existing methods for predicting cyberattacks do not fully take into account specific environmental factors of the target network, which may lead to deviation of the results from the real situation. In their work, the authors proposed a model for predicting cyberattacks based on the Bayesian network, formed on the representation of all vulnerabilities and possible attack paths, which, based on experiments, they claim gives more accurate results.

In 2005, an article by scientists E. Johansson, P. Johnson, "Assessment of enterprise information security - the importance of prioritization" [23], was distributed, committed to evaluating the level of data security in an venture. The paper considers the need of the data security sphere of the venture, specifically, which parts of data security are critical for the company and which are not.

A review of the forecasting methods used in cybersecurity is given in "Survey of Attack Projection, Prediction, and Forecasting in Cyber Security" [20] (2018) by M. Husák, J. Komarkova, E. Bou-Harb, P. Celeda. The following main tasks are discussed in the article:

1. Prediction of attack and recognition of intent, in which it is necessary to

predict the next step or intent of the attacker;

2. Intrusion prediction, in which it is necessary to predict upcoming cyber attacks;

3. Predicting a network security situation in which a cybersecurity situation is projected onto the entire network.

This paper also examines, compares and contrasts strategies based on discrete models such as assault charts, Bayesian systems and Markov models, as well as nonstop models such as time arrangement and Gray models. The authors talk about machine learning and information mining approaches that have as of late gotten a part of consideration and appear promising for an ever-changing environment like cybersecurity. The consider moreover centers on the commonsense utilize of strategies and their assessment. A comparative examination of strategies on discrete models (assignment charts, Bayesian systems and Markov models) is displayed, in comparison with nonstop models (time arrangement and Gray models).

An interesting model for the conceptualization of safety-related stress (SRS) is presented in J. D'Arcya, P. Teh "Predicting employee information security policy compliance on a daily basis: The interplay of security-related stress, emotions, and neutralization" [10] (2019). The authors proposed a theoretical model that connects SRS, discrete feelings, self-control reactions, and compliance with the ISP. The authors used an experiential sample design in which they interviewed 138 specialists. They noted that SRS has a positive relationship with anger and exhaustion, and they equate these negative emotions with neutralizing ISP disorders. Additionally, by reducing ISP compliance, frustration, and fatigue make employees more likely to follow the excuses of their ISP violations. Their results suggest that the neutralization of ISP disorders is not entirely stable, but may vary from one point to another in individuals. Therefore, according to the authors, when constructing the forecasting model, one should not forget about the human factor, or rather about the emotions and fatigue of workers.

Based on information from published articles in the field of security analysis and management, it can be concluded that an effective security management mechanism is still under development, however, correctly selected and formed conceptual models of security analysis and management, as well as modern machine

learning methods and algorithms, can give a more accurate predictive assessment of the level of enterprise security and suggest ways to improve it.

2. Preliminaries

2.1 Information security

Information protection is information security and support infrastructure from naturally or artificially unintentional or malicious effects, burdened with harm to information owners or users, and support infrastructure. Data security isn't limited to privacy safety alone. The subject of information relationships may suffer (suffering losses) not only from unauthorized access but also from system failure which caused a customer service break. Moreover, for many open organizations (for example, training), the protection of information itself does not come first.

To master the fundamentals of ensuring information security, a practical infrastructure has to be owned. Disclosure of certain main words is not an end in itself; initial ideas on information security priorities and strategies need to be developed.

Information protection is interpreted to mean a state that prevents the risk of accessing, modifying or damaging information by people not allowed to do so, as well as data spillage due to spurious electromagnetic radiation and interruption, uncommon capture attempts (destruction) amid transmission between computers.

Information protection is a set of measures aimed at ensuring the confidentiality and integrity of the processed and used information, as well as the accessibility of information for users.

Confidentiality - keeping secret critical information, access to which is limited to a narrow circle of users (individuals or organizations).

Integrity is a property, in the presence of which information preserves a predetermined appearance and quality.

Accessibility is such a state of information when it is in the form, place and time that the user needs, and at the time when he needs it.

The purpose of the protection and safety of information is to minimize management losses in a particular enterprise caused by a breach of data integrity, confidentiality or the inaccessibility of consumer information [9].

2.2 Web application firewalls

Protecting web applications is becoming critical to business security. Known security features such as firewalls and IDS/IPS can no longer guarantee security for public applications. Detection of anomalies at the network level, emphasis on signature analysis, outdated analysis and filtering techniques - these are just an incomplete list of reasons why standard perimeter security tools have become unsafe. In addition, such systems have too broad capabilities, which are reduced to controlling the security of the entire corporate network. This complicates their administration, which they constantly need.

Obviously, web applications require the installation of additional security features that specialize specifically in web technologies. A variety of technology companies have invested in the production of Web Application Firewall (WAF). For the first time, the task of automated filtering of queries that violate the application logic has reached the industrial level [41].

Each manufacturer has implemented algorithms, which in his opinion are a panacea for the security of web applications. Almost any solution declares its uniqueness and unconditional effectiveness. Different ideologies and their different implementation complicate the process of choosing the right solution, often leaving end customers confused. This happens due to diverse associations that arise among customers, even in terms of the key functionality of such a remedy. First of all, the question arises of what WAF should be like as a tool for protecting a real web application. Let's make a list of defense mechanisms that are inherent in WAF [41]:

1. protocol verification;
2. machine-learning reference model;
3. signature analysis;
4. specialized mechanisms;

5. user rules for detecting illegitimate requests;

6. denial of service attacks protection;

7. integration with third-party solutions.

2.2.1 Protocol verification

Protocol verification is the basic mechanism of passive protection against potential attacks that are carried out using atypical use of HTTP capabilities. Its goal is to leave the attackers as little room for maneuver as possible, limiting the request to special checks.

First of all, it is checking HTTP headers for RFC compliance, but this is not enough, and manufacturers go further, using restrictions on "best practices" and their own rules formulated in the process of analyzing possible vulnerabilities [41]. Typically, the following restrictions apply:

1. RFC requirements;

2. length and number of headers, parameters;

3. time frame;

4. validation of JSON, XML entities;

5. lack of invalid values.

2.2.2 Machine learning

This protection mechanism is key, and it is on it that the manufacturers of Web Application Firewall make the main bet in the development and promotion. Machine learning of the reference model is the process of adding web application access identifiers to a special model, followed by comparison of incoming requests to it. Matching queries with a learned reference model helps prevent both known and unknown vulnerabilities [41].

Machine learning-based protection mechanisms differ in:

1. data at the input of the algorithm;

2. the learning algorithm;
3. the way of decision making;
4. optimization technology;
5. configuration options;
6. the format of the reference models.

Characteristics of machine learning

1. *Input data for the training algorithm.* In addition to the access identifiers from the user's request, various implementations allow you to enter additional information that increases the effectiveness of training. For example, the following data may be considered [41]:
   (a) requests from trusted nodes (testing zone);
   (b) parameter type: dynamic, static, hidden, read-only;
   (c) web application responses.
2. *Learning algorithm.* The heart of machine learning, a mathematical device aimed at detecting anomalies in the deepest understanding of application software. For example,
   (a) Logistic regression classifier;
   (b) Support vector machine;
   (c) Multinomial Naive Bayes;
   (d) Random forest;
   (e) AdaBoost;
   (f) Artificial neural network;
   (g) Bag-of-words;
   (h) Principle component analysis;
   (i) Decision Trees.

3. *The way of decision making.* To avoid recursive learning, machine learning should apply clear criteria for the element's readiness to be included in the reference model. For example, these might be the following criteria [41]:

   (a) training time;

   (b) number of queries containing the trainable element;

   (c) thresholds for the element's entropy.

4. *Model optimization techniques.* During the development cycle, the protected web application constantly changes its behavior. Mismatch of the learned reference model with the real one leads to blocking of client requests. To prevent this from happening, each machine learning engine is equipped with techniques to optimize the reference model. The following techniques can be used [41]:

   (a) Interface for manual correction of the model;

   (b) Engaging a "teacher" for machine learning;

   (c) Heuristic analysis of queries that violate the current model, followed by its optimization.

5. *Configuration options.* Machine learning is a complex aggregate process, the need for configuration of which can depend on both quantitative indicators of traffic and qualitative characteristics of a web application. For example, you may need to configure [41]:

   (a) maximum training time;

   (b) operating threshold;

   (c) the way of decision making;

   (d) mathematical parameters of the learning algorithm.

6. *Reference Model Format.* Machine learning can be designed to build a model [41]:

   (a) Positive;

   (b) Negative.

Also, depending on the training, the model may contain various objects, in a typical implementation it is optimally considered to contain [41]:

(a) Resource Identifier (URI/URL);

(b) Application entity parameters (HTTP, XML/JSON entity);

(c) Session Identifier (Cookie).

2.2.3 Signature Analysis

Signature analysis is one of the oldest and most popular application security technologies. This is due to the fact that attacks on web applications in most cases are based on already known vulnerabilities using off-the-shelf tools. Moreover, the intensity of such attacks on the open Internet is so great that public web applications are exposed to them almost every minute [41].

In theory, the defense mechanism based on machine learning and constituting the reference model overlaps the need for signature analysis. Some WAF manufacturers are therefore abandoning the concept of signature developments. However, there are a number of situations where this protection mechanism is extremely useful - for example, during the training of the reference model, signatures are an additional line of defense against exploiting known vulnerabilities. In addition, requests that were deemed illegitimate are not submitted to the machine learning process, which makes learning more clean.

2.2.4 Specialized Protection Mechanisms

The effectiveness of filtering protective equipment substantially depends on the degree of overlapping of the same threats by various mechanisms. This is explained by the fact that there is always some probability within which the investigated attacks may go unnoticed for a specific defense algorithm. The probability of detecting an attack, if explored by many different algorithms, is sharply reduced. This is clearly demonstrated when trying to theoretically design a bypass of WAF security mechanisms [41] .

2.2.5 Custom rules for detecting illegitimate requests

WAF as an imposed security tool has great potential for analyzing requests passing through it. Here is just an incomplete list of internal processes through which any request passes [41]:

1. Decryption;
2. Inspection;
3. Inspection;
4. Parsing;
5. Normalization and storage;
6. Session management;
7. Security policy check.

The results of these processes should find application not only in the established protection mechanisms, but also be provided to IS administrators to form their own security rules [41].

2.2.6 Protection from DDoS-attacks

Ensuring the availability of information is as important as ensuring the confidentiality and integrity of the information processed by the web application. Despite the prejudice that denial of service attacks in network security should be prevented at lower levels than the application. WAF, as an application-level operator, offers interesting countermeasures.

2.2.7 Integration in the landscape of information security

All of these remedies will be related to each other. The web application firewall has extensive integration capabilities with other secure security systems. Today WAF can be integrated with system and service standards:

1. Vulnerabilities scanners;

2. Security information and event management;

3. Reputation service;

4. Fraud Prevention service.

This feature allows you to use the "Virtual Update" function, which allows you to automate control over the security of web applications. The scanner will find vulnerabilities, generate a report, and WAF on its basis forms open security rules. A key security aspect of many large companies is incident control systems. This makes it possible to correlate web application security events with events of other systems [41].

2.3 Machine learning Classifiers

We give the definition of ML used in this study. All ML models are regarded as classifiers and learn a function.

$$f(x) \mapsto y$$

An input point or example $x \in X$ is made up of n components or features and $y \in Y$ (e.g., malware or benign). In problems, the possible values of y are discrete. The output from the model represents real probability values for possible labels. In other words, there is a basic and almost always unknown distribution $D_{real}^{C_i}$ for each class C_i. The training data set X is sampled from this distribution, and the classifer approximates this distribution during training, thereby learning $D_{train}^{C_i}$. It is assumed that the test data set X_t, used to verify the effectiveness of the classifier is taken from the same $D_{real}^{C_i}$ [15].

Next, we present machine learning models for future use for our data under consideration.

2.3.1 Logistic regression classifier

The logistic regression model arises from the desire to model the posterior probabilities of K classes using linear functions with respect to x, while at the same time ensuring that they are summed with unity and remain in $[0, 1]$. The model

has the form

$$\log \frac{Pr(G=1|X=x)}{Pr(G=K|X=x)} = \beta_{10} + \beta_1^T x$$
$$\log \frac{Pr(G=2|X=x)}{Pr(G=K|X=x)} = \beta_{20} + \beta_2^T x$$

$$\log \frac{Pr(G=K-1|X=x)}{Pr(G=K|X=x)} = \beta_{(K-1)0} + \beta_{K-1}^T x \tag{2.1}$$

The model is specified with respect to the logarithmic coefficients $K-1$ or logistic transformations (reflecting the limitation that probabilities are summed up with unity). Even though the model uses the above class as the denominator in terms of probability, the denominator's choice is arbitrary in the sense that the grades are equal to that. A simple reckoning shows that

$$\begin{aligned} \Pr(G=k|X=x) &= \frac{exp(\beta_{k0} + \beta_k^T x)}{1 + \sum_{l=1}^{K-1} \exp(\beta_{l0} + \beta_l^T x)}, k = 1, \cdots, K-1, \\ \Pr(G=k|X=x) &= \frac{1}{1 + \sum_{l=1}^{K-1} \exp(\beta_{l0} + \beta_l^T x)}, \end{aligned} \tag{2.2}$$

and they are clearly reduced to one. To emphasize the dependence on the entire set of parameters $\theta = \{\beta_{10}, \beta_1^T, ..., \beta_{(K-1)0}, \beta_{K-1}^T\}$, we denote probabilities $PrPr(G=k|X=x) = p_k(x;\theta)$.

This model is simple if $K=2$ since there is only one linear function. It is used in biostatistical applications where binary answers are quite common (two classes). [16].

2.3.2 Support vector machine

The support vector machine (SVM) is a powerful classifier, which is recognized as a good choice of model for fitting multidimensional data. The basic theory of fitting a dividing line in a feature space is to maximize the gap between the nearest points to the line of each class. An example is shown in Figure 2.1, where the maximum field is the distance between the dashed lines and the red line.

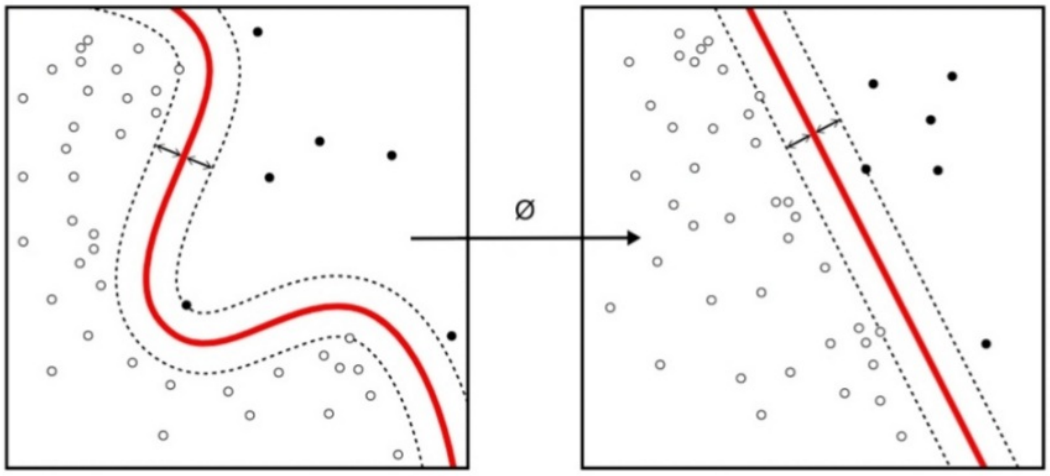

Figure 2.1: Two examples of SVM for a two-dimensional data set with their maximum fields

The main important argument of the SVM classifier is which kernel function to use. This core function dramatically changes the properties of the dividing line. Figure 4 shows how a dataset that is not linearly shared by a linear core can be classified much better using the radial basis function as the kernal.

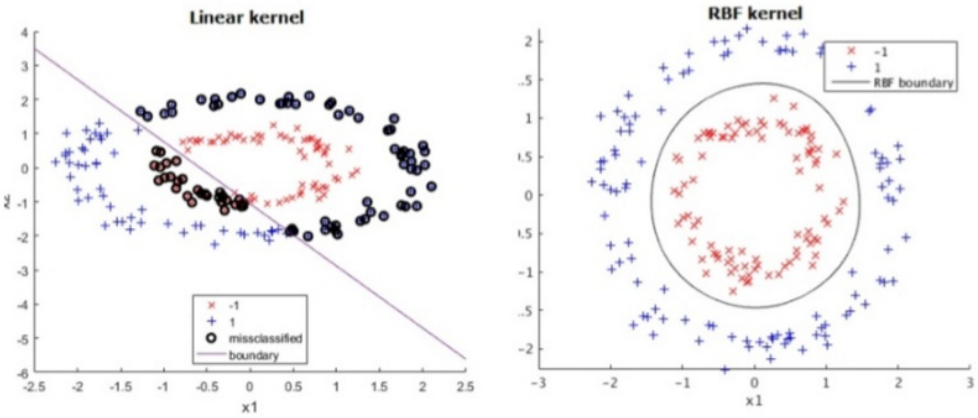

Figure 2.2: An example of how a radial basis function kernel can dominate over a linear kernel

We initiated by applying SVM to linearly shared data by constructing a matrix H from the point product of our input variables [12]:

$$H_{ij} = y_i y_j k(x_i, x_j) = x_i \cdot x_j = x_i^T x_j \tag{2.3}$$

$k(x_i, y_j)$ is an example of a family of functions called *Kernel Functions*. The kernel feature set consists of options (2.4) in the sense that they are all based on the calculation of the internal products of two vectors. This means that if functions can be transformed into space with a dimension using some potentially non-linear function of displaying features $x \longmapsto \phi(x)$, then only the internal products of the displayed inputs in the feature space should be determined without the need for explicit calculation of ϕ.

The reason that this *Kernel Trick* is useful is that there are many classification/regression problems that are not linearly separable/regressive in the input space x, which may be in the attribute space with a higher dimension with a suitable mapping $x \longmapsto \phi(x)$.

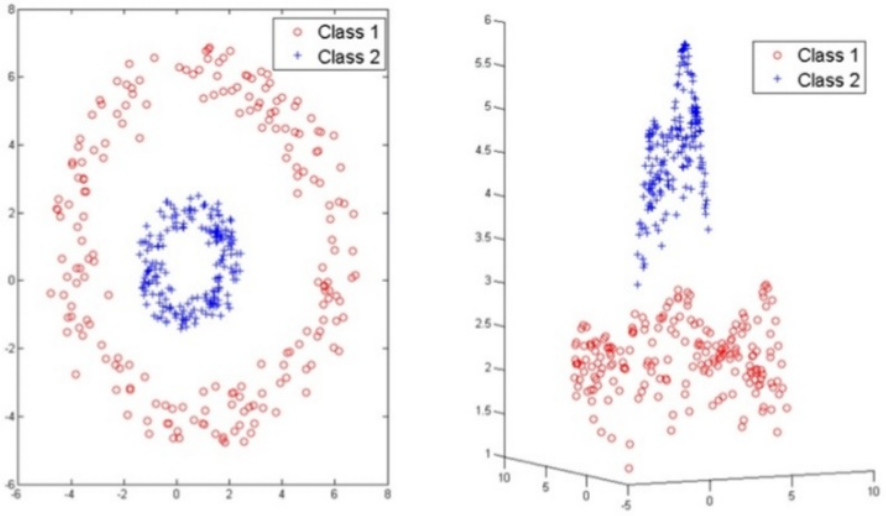

Figure 2.3: Dichotomous data reassigned using Radial Basis Kernel

Referring to Figure 2.3, if we define our kernel to be:

$$k(x_i, x_j) = \epsilon^{-(\frac{\|x_i - x_j\|^2}{2\sigma^2})} \tag{2.4}$$

then the data set, which can not be divided linearly in the two-dimensional data space x (as in the left part of *Figure 2.3*) is separated in the nonlinear feature space (right side of *Figure 2.3*), a function implicitly defined by this nonlinear core - known as *a Radial Basis Kernel* [12].

2.3.3 Multinomial Naive Bayes

The Bayesian classifier uses the Bayesian rule shown in Equation 2.5 to assign the probability that the data point is in a particular class. You can then determine the threshold of how likely the result of equation 2.5 should be in order to classify the input point as a positive class.

$$p(C_i|\vec{F}) = \frac{p(\vec{F}|C_i)p(C_i)}{\sum_j p(\vec{F}|C_j)P(C_j)} \tag{2.5}$$

where $p(C_i|\vec{F})$ - probability of a data point with features $\vec{F}$ belonging to class C_i; $p(C_i)$ - the probability of class C_i occuring in the data set, i.e. our prior knowledge about the data set; $\sum_j p(\vec{F}|C_j)P(C_j)$ - a normalizing factor to ensure that $p(C_i|\vec{F})$ is a propely defined density.

The naive part of the classifier is that $p(\vec{F}|C_i) = \prod_j p(F_j|C_i)$, that is the probability of the feature vector can be represented as an independent multiplication of the probability of each feature. Bayesian classifiers are very popular among text classification tasks, and therefore the naive part assumes that the words of the text are independent of each other. This clearly does not apply to the text, but still works very well. A polynomial naive Bayes classifier also takes into account the number of occurrences of a particular word in the input string [30].

2.3.4 Random forest

Bagging or bootstrapping is a method of reducing expected variance in the forecast function. Hashing works perfect for large deviation procedures such as trees. For regression, we simply fit the same regression tree many times to train data versions with initial loading, and thus average the output. Each vote for the predicted class is by the tree committee, for classification. Random forests are a major alteration of the bags, creating and averaging a large collection of decorrelated trees.

The output of random forests on several possible things is very close to growing, making them easier to train and to tune. As a result, random forests are popular in different packages and any noisy, but roughly impartial, models are implemented and therefore the spread reduces.

Trees are ideal for packaging in packages, because when grown deep enough,

they can capture complex data interaction structures and have a relatively low bias. The trees are extremely noisy, so they profit greatly from averaging. Since the tree generated in the package is distributed uniformly (i.e.), the expectation of the average value of such trees corresponds with the expectations of every one of them. This means that the displacement of the trees in the bags is identical to that of the individual trees (initial loading). [16].

Consider a random forest algorithm for regression or classification given in [16]:

Algorithm Random Forest for Regression or Classification:.

1. For $b = 1$ to B:

   (a) Draw a bootstrap sample Z^* of size N from the training data.

   (b) Grow a random-forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached.

      i. Select m variables at random from the p variables;

      ii. Pick the best variable/split-point among the m;

      iii. Split the node into two daughter nodes.

2. Output the ensemble of trees $\{T_b\}_1^B$

To make a prediction at a new point x:
Regression: $f_{rf}^B(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$.
Classification: Let $C_b(x)$ be the class prediction of the bth random-forest tree. Then $C_{rf}^B(x)$= majority vote $\{C_b(x)\}_1^B$ [16].

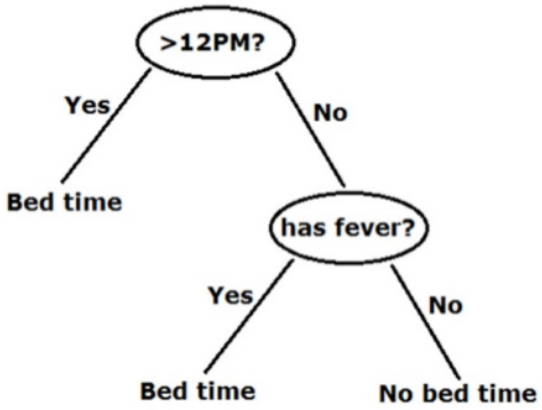

Figure 2.4: An example of a decision tree

To describe the random forest classifier, you must first define a decision tree. In fact, the decision tree is a series of yes/no questions, like the questions in Figure 2.4. The nodes contain questions, answers to questions. The random forest speaks for itself; it consists of several randomly generated decision trees. Solutions for all trees will be classified as random forests.

2.3.5 AdaBoost

Like random forests, the AdaBoost classifier combines several classifiers. AdaBoost is best used when combining weak classifiers, such as decision trees of only height 1. The algorithm starts with training a model based on training data and then creating a second model to correct errors from the first model. This additive procedure is carried out until the maximum number of models is created or until the ideal forecast is reached.

The boosting algorithm AdaBoost [32]

Given: $(x_1, y_1), ..., (x_m, y_m)$ where $x_i \in \chi, y_i \in \{-1, +1\}$.

Initialize: $D_1(i) = 1/m$ for $i = 1, ..., m$.

For t=1,..., T:

1. Train weak learner using distribution D_t.
2. Get weak hypothesis $h_t : \chi \to \{-1, +1\}$.
3. Aim: select ht with low weighted error: $\varepsilon_t = Pr_{i \sim D_t}[\, h_t(x_i) \neq y_i]$.

4. Choose $\alpha_i = \frac{1}{2} ln(\frac{1-\varepsilon_t}{\varepsilon_t})$.

5. Update, for $i = 1, ..., m$: $D_t + 1(i) = \frac{D_t(i) exp(-\alpha_t y_i h_t(x_i))}{Z_t}$, where Z_t is a normalization factor (chosen so that D_{t+1} will be a distribution).

Output the final hypothesis: $H(x) = sign(\sum_{t=1}^{T} \alpha_t h_t(x))$.

2.3.6 Artificial neural network

In recent years, the popularity of artificial neural networks (ANNs) has expanded significantly. The thing is that it can be any function that allows you to match even the most complex data.

The ANN type discussed in this thesis has the simplest structure, a direct-connected neural network, shown as an example in Figure 2.5. This particular example has three inputs, one hidden layer with three "neurons" and two outputs, but the fact is, that the network can consist of any number of layers and neurons. As Figure 2.5 can explain, the input of a neuron in layer l is a linear combination of the outputs of neurons in layer $l-1$. Each of these inputs in each layer is weighted, and the weighted combination of inputs into a neuron is converted to zero or one, depending on the function of activation of the neuron. The determination of all weight matrices $W^{(l)}$ of each layer l is a learning process. Studying weights requires a labeled dataset and is performed using the backpropagation algorithm, but this is beyond the scope of this report [17].

2.3.7 Bag-of-words

Since machine learning classifiers always need numbers as input, raw text data is not suitable for direct use in a model. A popular method for solving the problem of entering textual data is called a word bag, which converts a text string into a word quantity vector. Converting a word bag leads to the fact that each unique word in the entire data set is represented as its own feature. The vector of data point objects is simply zero for each object, except for objects representing words that exist in the string. The values of these functions are equal to the number of occurrences of words in a string. An example of conversion using a word package is shown in Figure 2.6, where the data set contains two lines, and the attribute space is each unique word in the data set [40].

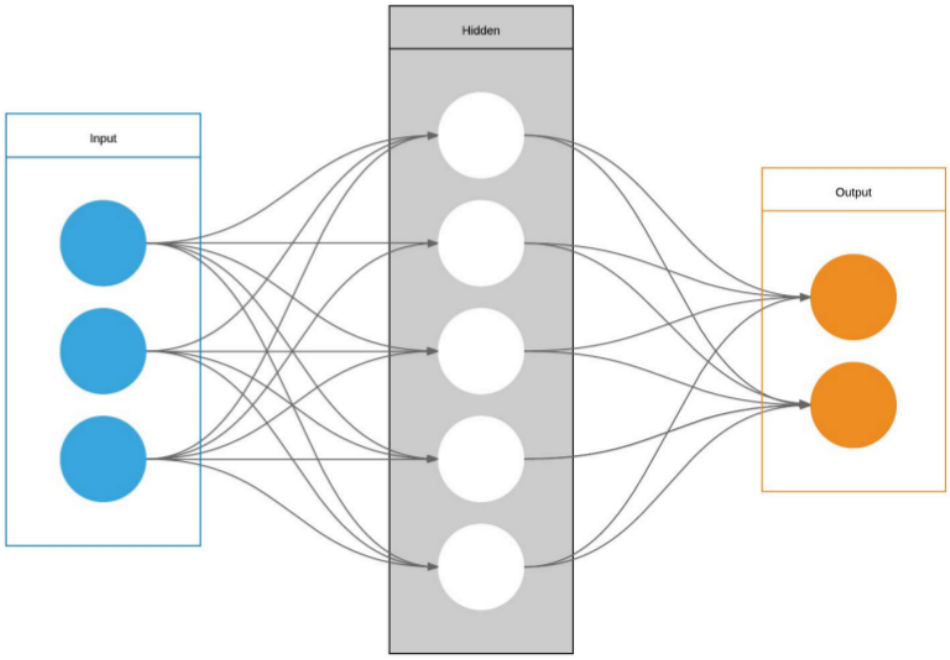

Figure 2.5: An example of an artificial neural network

2.3.8 Principle component analysis

The "most suitable" line can be expressed as a line that minimizes the average square distance from one point to a line, given a collection of points in two- or more-dimensional space. Furthermore, the most appropriate line can be chosen from the directions perpendicular to the first in a similar manner. Repeating this process provides an orthogonal basis on which there is no correlation between different individual data dimensions. These main vectors are referred to as core components, as well as a core component analysis of many associated procedures.

The principle of component analysis (PCA) falls into the category of teacher-less learning algorithms. This is a dimensional reduction method that projects a data set from the original space of objects into a new reduced space of objects. New features are linear combinations of original features. The purpose of the conversion process is to maximize dispersion retention. PCA can be used to compress object spaces, select objects, and render. We use PCA to visualize large data.

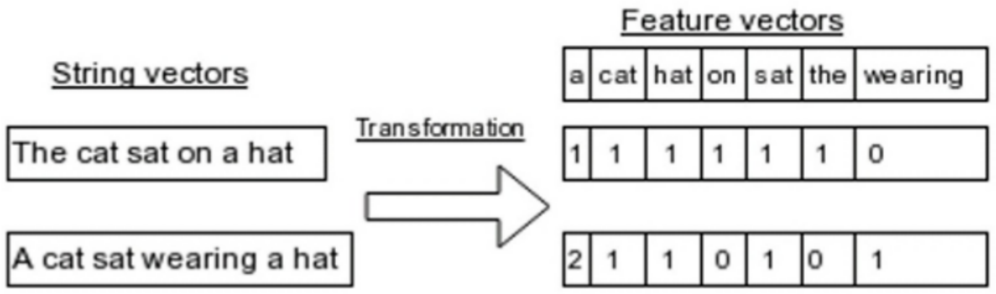

Figure 2.6: Example of two lines converted to a common feature space

A PCA is defined as an orthogonal linear transformation that converts data into a new coordinate system, where the largest variance over a projection of scalar data is in the first coordinate, the second largest variance in the second coordinate, and so on.

Consider an $n \times p$ data matrix, $\mathbf{X}$, with an empirical average zero in the column (the average sample value of each column is offset to zero), where each of the rows n represents a different repetition of the experiment, and each of the columns p gives a special feature.

Mathematically, the transformation is determined by a set of sizes $\mathbf{l}$ of p-dimensional vectors of weights or coefficients $w_{(k)} = (w_1, ..., w_p)_{(k)}$ that map each row vector x_i of X to a new vector of principal component *scores* $t_{(i)} = (t_1, ..., t_i)_{(i)}$, given by

$$t_{k(i)} = x_{(i)} \cdot w_{(i)} \quad for \quad i = 1, ..., n \quad k = 1, ..., l \tag{2.6}$$

The individual variables $t_1, ..., t_l$ of t examined over the data set successively inherit the maximum possible variance from X, with each coefficient vector w bounded to *unit vector* (where l is usually chosen to be less than p to reduce the dimension).

2.3.9 Decision Trees

The decision tree is a simple presentation for examples. All objects are assumed to be finite objects called "classification". Each item in the classification area is called a class. A decision tree or classification tree is a tree in which each internal

(not final) node is marked with an input function. The arcs that go from the node marked with the input object are marked with each of the possible values of the target or output object, or the arc leads to the subordinate decision node on the other input object. Each leaf of the tree has classical or distributed probabilities by classes, which means that the data set was classified either by specific classes or by specific capabilities (which, if the decision tree is well-built, are inclined to certain subset classes).

The tree is built on the basis of a standard source set, which contains successor descendants. The basis on a set of separation rules, the basis on the basis of classification. This process is repeated for each derived subset in a recursive way, called a recursive standard. The recursion is completed when all values have the same values or when splitting does not matter for the forecast. This process is evidence that the most common strategy is to study trees based on data.

Data can be used as a combination of mathematical and computational methods that can help describe, categorize, and generalize data in a dataset.

The data comes in the form:

$$x, Y = (x_1, x_2, x_3, ..., x_k, Y) \tag{2.7}$$

The dependent variable, Y is the target variable that we are trying to understand, classify or generalize. The vector x is composed of the features, x_1, x_2, x_3 etc., that are used for that task.

3. Analysis of methods

This chapter provides analysis of methods for ensuring the security of information systems in the enterprise.

3.1 Security Risk Classification

In studies [34] considered a structured approach to identify groups of threats specific to an enterprise, which is an important step for security planners involved in developing cost-effective strategies to address their organizations' information security risks. The performance of the information security risk management plan of the organization is focused on the effective detection of threats to the information systems of the organization. In [34], a set of five categories of high-level threats is presented:

1. Staff and administration;
2. Networks;
3. Hardware;
4. Software;
5. Environmental and physical safety.

Within these categories, 21 factors were identified, which were shown in Table 3.1 by [34]. Approximately 450 recommended security measures have been categorized into these threat categories.

However, the work with five types of threats has been troubling. Although a quasi-analytical approach was taken to assess the interpretation of the system's security situation, the study became ever more subjective. Several anomalies

Table 3.1: Legacy Threat List

Threat Category	Threat
Personal/ Administrative Threat	Terroist Actions/Civil Disorder Activity for Personal Gain Malicious Acts by an Individual Employee Tampering with or Destruction of Hardware and/or Related Components Theft of Hardware and/or Related Components Theft of Resources
Network Threat	Essential Communication Line/Equipment Failure Masquerading as an Authorized User Spoofing Wiretapping or Eavesdropping
Hardware Threat	Essential Hardware Failure
Software Threat	Programmer/Operator Error Essential Software Failure Malicious Software Invasion Unauthorized access or Execution Privileges
Environmental/ Physical Security Threat	Theft or Equipment Tampering Loss of stable Electrical Power Facility or Equipment Fire Natural Disaster Temperature/Humidity Extremes

that could cast doubt on the reliability of the overall results were noticed when considering the list of threats.

A comprehensive and balanced list of threats must, therefore, be drawn up from which to protect information systems. This article outlines the method in the form of a phrase used to compile a list of threats to enterprises.

A security situation and additional security measures are subjectively determined to reduce the risk, based on the five threat categories presented [34].

One of the organization/enterprise's tasks is a thorough and systematic knowledge of the risks to their information assets and how to procure the necessary funds to remove them, which could continue to secretly generate an information leak. In order to better understand security threats, a model for classifying security threats is proposed in [24], which allows us to study the influence of a class of threats rather than the impact of a threat as the threat changes over time and takes into account various criteria for classifying the security risks of information

systems and provides an overview of the classification models for most threats as well.

In research [24] divided the approaches to the classification of threats into two main classes:

1. Classification methods based on attack methods
    (a) The three-dimensional orthogonal model - a threat model for classifying security threats that solves the problem by introducing a three-dimensional model that divides the threat space into subspaces according to the three orthogonal dimensions designated as motivation, location, and agent.
    (b) Hybrid model for threat classification
        i. Frequency of security threats;
        ii. Scope of security threats;
        iii. Source of security risk;
    (c) Information Security Threat Classification Pyramid Model - Classifies intentional threats based on three factors:
        i. The knowledge of the attackers about the system;
        ii. The criticality of the area;
        iii. Losses that may occur in the system or in the organization.
2. Threat-Based Classification Methods
    (a) **STRIDE Model** (Spoofing identity, Tampering with data, Repudiation, Information disclosure, Denial of service and Elevation of privilege) - this is a focused approach, in which an attempt is made to penetrate the consciousness of an attacker by assessing threats.
    (b) **ISO model** - The ISO standard (ISO/IEC 27001) in order to formulate comprehensive information security requirements defines three main indicators [21]:
        i. assessment of the risks faced by the organization (identification of threats to resources, their vulnerability and probability of occurrence of threats, as well as possible damage);

ii. compliance with legislative, regulatory and contractual requirements that must be met by the organization itself, its business partners, contractors and service providers;

iii. formation of a set of information processing principles, objectives, and requirements developed by an organization to support its activities [21].

Most security threat classifications are generally restricted to the use of one or two threat classification criteria (not all threats are included in the classification), and their definitions are not mutually exclusive. This can suffice for a stable environment (a small organization), where security threats are fairly constant but organizations can not defend against internal threats in an ever-changing environment [14].

In fact, organizations are exposed to several types of threats that affect their reputation and to reduce their risks, it is important that they identify all the characteristics of the threats.

Classification allows an organization to know the threats that affect their assets and areas that may be affected by each threat, and thus to protect their assets beforehand. It also helps managers build their organizations' information systems with a lesser degree of vulnerability [14]. In the work of existing threats, the principal problems can be identified. Present classifications do not, in fact, follow the concepts of classification [26], [37], [18]. The usual approach at this point is to combine various classifications, and to create a hybrid. Because of the above results, [24] proposes a hybrid model for the classification of threats to the security of an information system, which has been called a multidimensional model for the classification of threats in order to comply with all threat classification principles.

In the next paper [28], "threats to the safe use of the Internet" were assessed by interviewing voters asking the following questions: "Which threat struck you the most?", "What do you think, which threat has a significant impact on society?" etc., dividing the respondents into three groups: "organizations", "users" and "system administrators/developers", and identified 10 major security threats. Appropriate threats were assigned to each group, and then information was collected, including a summary of the incident, how it occurred, the degree of damage and how it was caused, and what measures were taken. The following threats were presented in three groups:

1. Threats to organizations:
   (a) DNS cache poisoning threat;
   (b) complex target attacks;
   (c) daily information leakage.
2. Threats to users:
   (a) various ways of infection of computer viruses and bots;
   (b) threats due to wireless LAN encryption vulnerability;
   (c) never reduce spam;
   (d) threats arising from the use of the same user ID and password.
3. Threats to system administrators/developers:
   (a) Threats of attacks through a legitimate site;
   (b) Actual passive attacks;
   (c) Potential vulnerability of embedded systems/devices.

Next, we consider cloud computing in the corporate infrastructure, discussed in [4], which addresses the threats and security problems in cloud computing and the enlightened steps that an organization may take to the security risks and protect its resources.

Cloud computing is a model for providing convenient network access on demand for a common fund of configurable computing resources (for example, data networks, servers, storage devices, applications, and services - both together and separately), that can be delivered quickly and released at minimal operating costs or calls to the provider [22].

Enterprises are beginning to see cloud computing technology as a way to reduce costs and increase profitability, because in all sectors, CIOs constantly have to reduce capital assets, number of employees and support costs, and cloud systems enable them to achieve these goals. Figure 3.1 shows the available resources for enterprises in the cloud (Brandl, 2010) [4].

Cloud computing profitability can be explained in the form of the "cost associativity" as shown in formula (1). The left side multiplies the net gain per

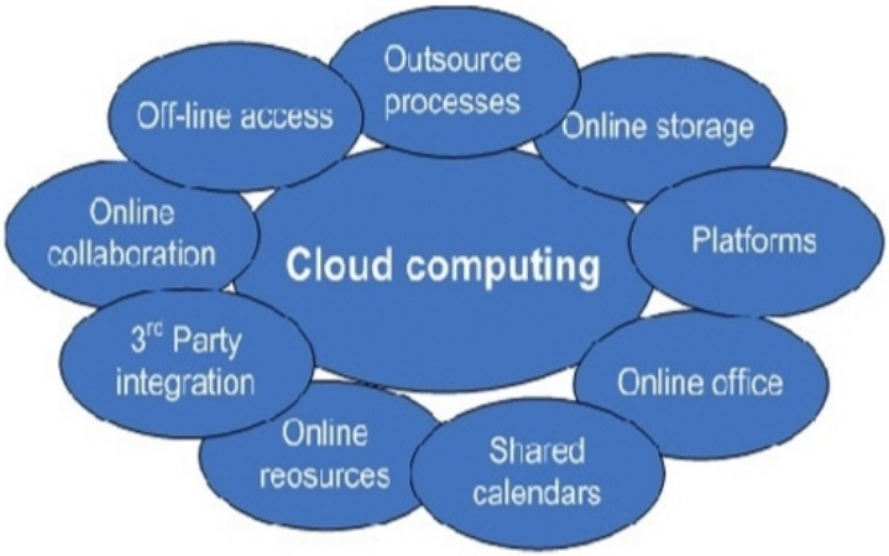

Figure 3.1: Cloud Computing Resources (CloudTweaks, 2010) [4]

usage hour by the number of user hours, generating the estimated benefit from the use of cloud storage, while the right side conducts the same equations for a fixed capacity data center, taking into account the actual load, including off-peak data center workloads; which side is higher, the more benefit opportunities [13].

In [13] (Armbrust et al., 2009, p. 10-11) gave an example of elasticity with calculations of potentials for saving cloud computing and reducing costs:

$$UserHours_{cloud} \times (revenue - Cost_{cloud}) \geq$$
$$\geq UserHours_{datacenter} \times (revenue - \frac{Cost_{datacenter}}{Utilization})$$

There are several large cloud computing providers, including Amazon [21], Google [22], Salesforce [23], Yahoo [24], Microsoft [25], Alibaba [26], IBM [27] and others who use the services Software as a Service (SaaS), Platform as a Service (PaaS), Storage as a Service and Infastructure as as Service (IaaS) software and infrastructure as a service.

Cloud computing is facing the same number of security threats currently being encountered in existing computing platforms and Internet enterprise networks. These threats and risk vulnerabilities come in many forms. The Cloud Security Alliance [3] conducted a study of the threats facing cloud computing and identified

seven key threats:

1. cloud abuse and misuse;
2. unsafe application programming interfaces;
3. evil insiders;
4. common technology vulnerabilities;
5. data loss/leak;
6. account, service and traffic theft;
7. unknown risk profile.

According to [29], the transfer of your data to a cloud service is like "putting all your eggs in one basket". Studies have shown that attackers can identify where the data is in the cloud, and use different tricks to gather information. It needs careful preparation and awareness of emerging risks, challenges, vulnerabilities, and potential countermeasures for the effective adoption of cloud computing at the enterprise. Before applying this technology, it is believed in [4] that a company should analyze the security risks, threats, and existing countermeasures of a company/organization.

3.2 Methods of detection

Detection of known web-based attacks is done using signature-based detection (SBD), whilst the detection of HTTP requests for the anomaly is done using anomaly-based detection (ABD). ABD based on learning is implemented using Artificial Neural Networks (ANN). Thus, learning-based ABD is assured by using ANN to adapt the model against zero-day attacks.

Detection of known network attacks is performed using signature-based discovery (SBD), while HTTP request discovery for anomaly is performed using anomaly-based detection (ABD). Learning-based ABD is implemented using artificial neural networks (ANNs). Thus, a training-based ABD is provided by ANN to adapt the model to zero-day attacks [38].

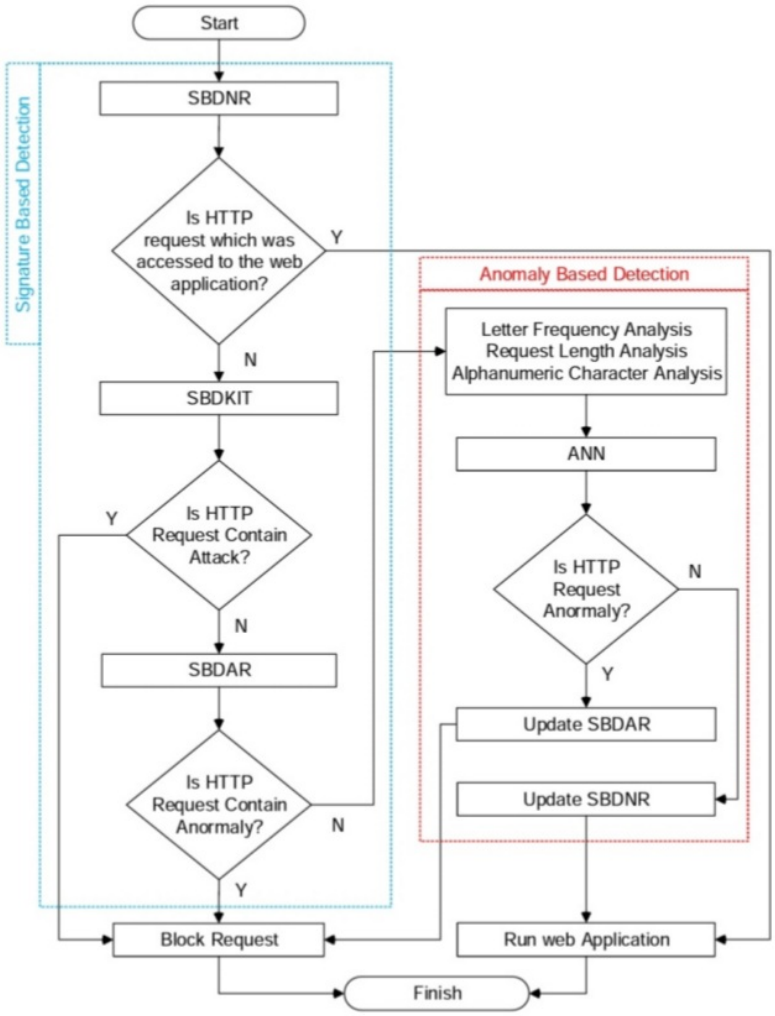

Figure 3.2: Model Flow Chart proposed [38]

Three stages of detection are required through SBD. The first step is SBDKIT, which is SQL injection, crossite scripting (XSS), command injection and directory attack. An SBDAR checklist that has been updated as an HTTP request. They can be blocked by SBDAR without using ABD. The final step in SBD is to update the SBDNR with HTTP requests that were previously defined as regular HTTP characters using ABD. HTTP records that can be detected using SBD are routed to ABD HTTP requests. These characters are alphanumeric character analysis. ABD was implemented using ANN. This is one of the methods of artificial intelligence that is used as input. HTTP voices coming to web applications. All you need is regular HTTP records that will easily access your web application without

using ABD. This feature will improve the detection rate of the performance of the proposed model. Therefore, in the HTTP request, there are web applications with updated SBDNR and SBDAR checklists, the disadvantage of SBD, which is ineffective in zero-day attacks, can be eliminated. The considered model of the technological scheme is presented in Figure 3.2 [38].

4. Basic criteria and requirements for information security system models

4.1 Web application structure and attack types

With the development of technology, each company has its own website. A website, in turn, is a system of electronic documents (data files and code) owned by an individual or organization, and can be accessed on a computer network under a common domain name and IP address, or locally on a single computer. In general, a web interface that faces the public Internet is considered to be the most vulnerable and "risky" when it comes to vulnerabilities, so websites are one of the hackers' main targets. When an intruder arrives at the web site, it may allow them to enter the settings of the computer system or database, server, or installation or operating system.

Here are some analyses of hacking sites at enterprises.

On average, between 30 000 and 50 000 sites are hacked daily, and in reality, most of these 30 000 sites are legitimate small businesses that send malicious code to cybercriminals unwittingly.

Based on currently available figures, 64% of businesses suffered web attacks. 62% of the companies were exposed to phishing and social engineering attacks and 59% of the companies had malicious code and botnets [36].

According to the above data, we see the importance of the need to ensure the information security of web applications of enterprises that daily use Internet resources.

In this regard, we will consider the structure of web applications and provide

types of attacks to investigate further work.

We will need to learn the components of the web application to grasp the vulnerability of the web environment. Web applications contain three main components, as shown in Figure 4.1. The programming language is used to build part of the client and to construct queries about databases. Hypertext Transfer Protocol (Http), which is used to connect client-side to server-side communication. A business process is also a distinctive part of any web application.

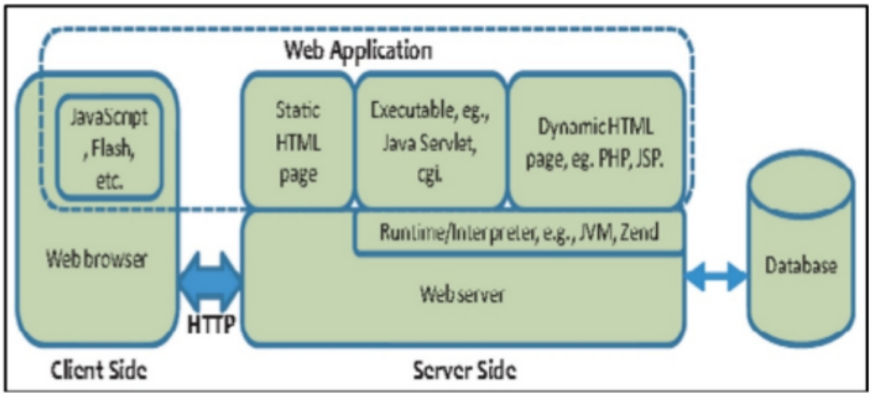

Figure 4.1: Web Application Components [1]

Another major threat to Internet security and web applications that users view through the default port number is 80 using the http protocol and port number is 443 using the https secure layer protocol. An attacker begins to use the Internet as a regular client or website user, then these ports are used to attack the site and access client data and files. The size of the attack depends on the importance of the data and business of the companies that own these sites [1].

The following are the most common types of attacks in web applications:

1. DDoS attacks. DDoS attacks are aimed at suppressing the target web application/website/server with false traffic, reducing network bandwidth and making it inaccessible to legitimate users. Some common but dangerous types of DDoS attacks include DNS enhancement, Ping of death, Smurf attacks, HTTP flood, SYN flood, etc. [35].

2. Structured Query Language (SQL) Injection is a code injection technique used to modify or retrieve data from SQL databases. In such attacks, the attacker injects malicious SQL code into user input fields in web applications in the form of requests or requests, such as submitting forms, contact

forms, etc. Thus, they gain access to the internal database of the application in which they infiltrate to extract confidential customer or company information, receive unauthorized administrative access, modify or delete data, etc., or even gain full control of the web application.

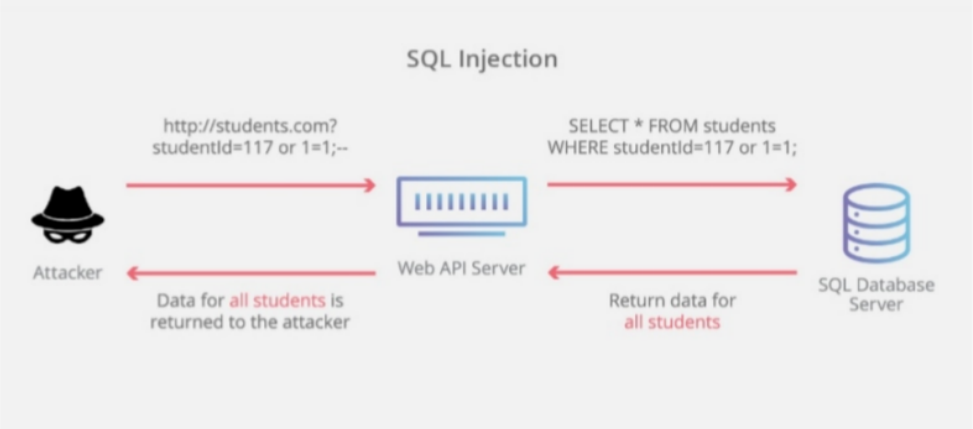

Figure 4.2: SQL injection [5]

3. Cross-site scripting attacks (XSS). XSS attacks target vulnerable web app/website users to gain access to and monitor browsers. Here, when an unsuspecting user loads the application/website, attackers use vulnerabilities and spaces in the application to inject malicious scripts/codes which are executed. XSS attacks compromise the personal and confidential information of the user and often lead to identity theft, hijacking of sessions, etc.

4. Zero-day Attacks. Zero-day attacks are those in which the organization only becomes aware of the existence of hardware/software vulnerabilities when the attack occurs. This is unexpected and therefore very harmful for the business because they have no easy fixes or improvements to secure their application [35].

5. Business Logic Attacks. Business logic is a critical element that connects and transmits information between the user interface and the databases and software systems, enabling users to use the web application/website efficiently. Where business logic has gaps, errors, or matches, this creates vulnerabilities that cyber attackers often use to gain monetary and other

benefits. Attackers don't use distorted requests and malicious load to organize business logic attacks. They use legitimate values and legal demands in the application to exploit indirect vulnerabilities [35].

6. Man-in-the-middle attacks. These attacks occur when cybercriminals position themselves as one of two parties between the application and authorized users to steal sensitive information, such as passwords, login credentials, credit card information, etc. An attack may be organized using simple means, such as providing free, malicious access points in non-password protected public places. When victims connect to these access points, they give the attacker complete transparency of data exchange on the Internet [35].

7. Malware. Malicious attacks are organized by exploiting vulnerabilities in the application or using social engineering methods such as phishing to inject malware to a website/web application/server, such as trojans, ransomware, spyware, rootkits, etc. Therefore an attacker gains access to confidential information, confidential parts of the application, changes in system configuration[35].

8. Defacements. Attackers modify the content of the website in attacks using falsification, the easiest of all cyber attacks, and replace it with their own content to reflect the political ideology/agenda, shock users with contentious messages or images. The Web application may become inaccessible to users until the corruption is fixed [35].

By examining the above works, we have identified what security threats exist in organizations, and reviewed the structures of web applications with common types of attacks.

Traditional protection solutions such as network firewalls, intrusion detection systems (IDS), and intrusion prevention systems (IPS) are good for stopping unwanted traffic and maintaining network-level protection. But they don't have the ability to detect and stop SQL injection, session hijacking, cross-site scripting (XSS), and other attacks that occur as a result of web application-inherent vulnerabilities. One solution is to configure the WAF (Web application firewall).

4.2 Justification Web application firewall and its advantages

Web application firewalls provide an efficient threat detection solution by testing incoming HTTP requests before entering the server. WAF (Web application firewall) detects and blocks malicious website traffic-related attacks that may have leaked via conventional security solutions. WAF is also used by organizations to help them fulfill HIPAA and PCI-DSS requirements[2].

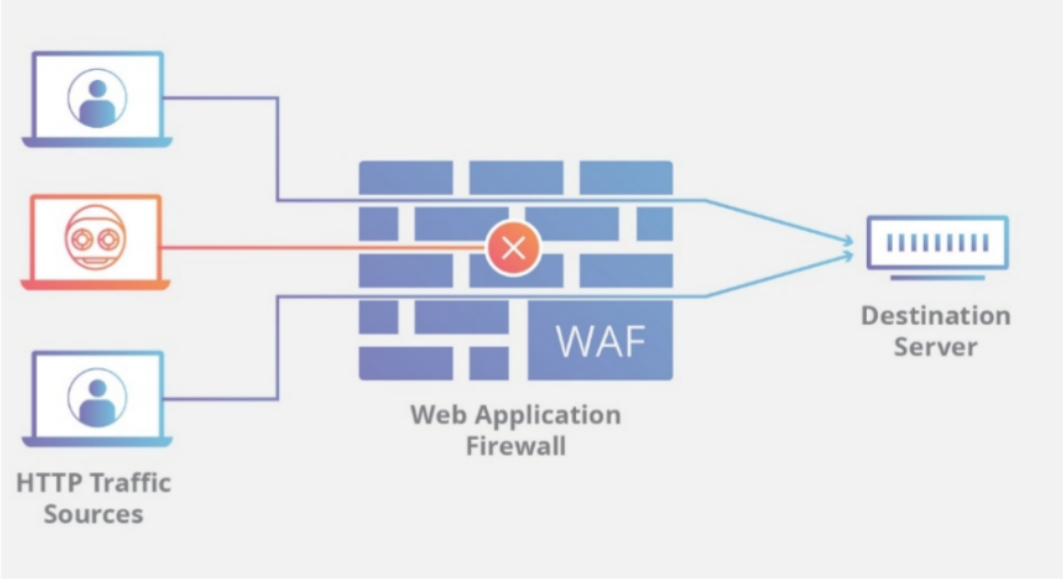

Figure 4.3: WAF protection structure [5]

WAF performance is typically based on one of three security models:

1. Blacklist or negative protection model-this uses common signatures to protect the site from known attacks and special signatures to prevent web application vulnerabilities that could be exploited;

2. Whitelist or positive model of safety-this uses signatures and sometimes additional logic to allow only traffic that fulfills some criteria. An example is allowing only HTTP GET requests from a specific URL and blocking everything else;

3. Hybrid safety model-this This applies to negative as well as positive models. Some of the configurable parameters include request blocking, session

blocking, blocking IP addresses, blocking users or logging out users [2].

WAF has some disadvantages due to the lack of automation, scalability, and coverage of emerging threats, as modern botnets become more and more efficient and aggressive. These botnets are now created using an artificial intelligence (AI) feature on top of the "old" Internet of things (IoT) botnets, which are becoming more versatile in their ability to attack with different vectors. The functionality that the classic WAF offers has become a subject of discontent, while next-generation WAFS, which were born as artificial intelligence systems that can handle such a multidimensional complexity of threats, are quite rare.

There are not many artificial intelligence/machine learning (AI/ML) solutions in the cyber security and application security segment. However, more and more AI and ML solutions are starting to show up as the main success against the distributed denial of service (DDoS) attacks and, more specifically, against the world of DDoS (Distributed Denial of Service) applications, which was demonstrated by L7 Defense with its uncontrolled approach to learning. Such technology can also play a decisive role in WAF solutions, such as protection against the same multipurpose botnets [8].

As for protection, traditional methods use some dictionary or database of known vulnerabilities.

The WAF should be able to fight a variety of multi-vector attacks, such as SQL injection, remote command execution, enabling remote files, enabling local files, implementing PHP, implementing LDAP, implementing Memcache, and cross-site scripting (XSS); all together. We need experience to identify these types of attacks and classify them with maximum accuracy from the first request. Objectively, this is a critical part for determining the "very first request» [8].

False positives and false negatives should be limited, which would be close to zero at the webpage level.

We need to find a solution that uses the same AI concepts from the Applicative DDoS and need to add specific classifications in addition to being able to algorithmically dynamically identify all types of attacks on the fly. The WAF should be given an additional capability so that it can accurately determine if there is some kind of traffic aggregation in your web interface coming to you, using machine learning, which will be our further work.

For high-quality organization of activities to prevent threats to information

security of the enterprise, it is necessary to:

1. high-quality threat forecasting;

2. development of preventive measures to protect the enterprise.

In the organization of activities to prevent threats to information security, the main task is to predict and evaluate possible threats to the enterprise.

Thus, it is very important that the enterprise adopts a structured methodology to identify relevant threats, reassess remaining vulnerabilities, and identify new threat.

5. Implementation of key indicators defining security models

5.1 Preprocessing of implementation

Using machine learning methods to detect intrusions and firewalls has become a highly researched area in recent years, so it was necessary to conduct a little preparatory work before starting any implementation by reading the relevant topics with the latest published results presented in the literature review. The workflow of this study is close to how usually machine learning projects are applied, or to a method of data analysis. The project is conditionally divided into five distinct organized stages: query or problem identification, data processing and data clearing, task creation, system preparation, and evaluation. Although the process appears to be very consistent, that was not always the case. Extracting functions, for example, helped us find new data information which required further data cleaning and even clarification of the initial question. Evaluation of various classifiers and functions has taken us back to the data and functions. Some of the inappropriately labeled data is obviously unacceptable, prompting a further round of cleaning of data. See Figure 5.1 for a schematic representation of the workflow.

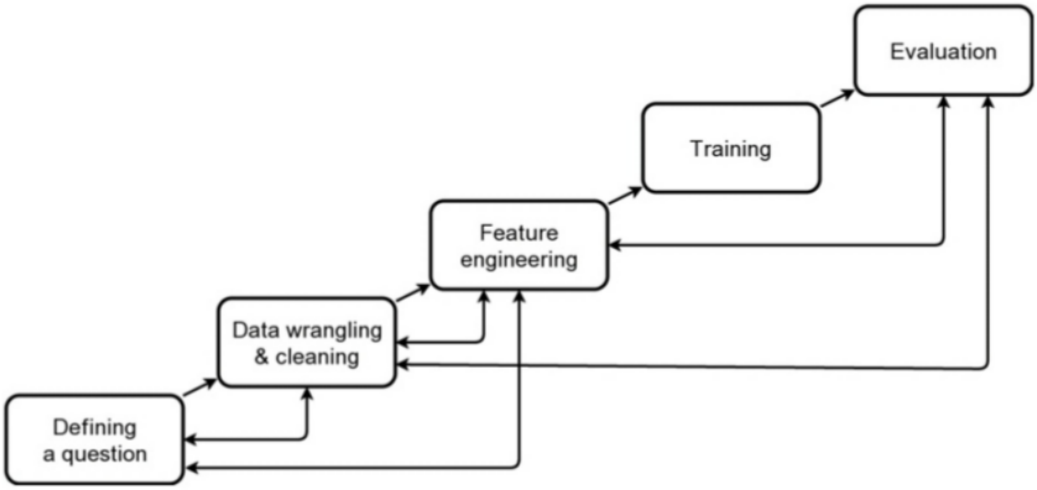

Figure 5.1: Visualization of the workflow.

The programming language is used by Python. Since the language is very suitable for data analysis, and thanks to the libraries described below, it is the most popular programming language for machine learning. These are the main libraries that are widely used in this thesis to significantly reduce the amount of code needed to write ourselves:

NumPy	Numerical Python, a library that adds support for large multidimensional matrices and a collection of high-level mathematical functions. Website: http://www.numpy.org/
Pandas	Build upon NumPy and provide high-performance table-like data structures well suited for handling and manipulating large sets of data. Website: http://pandas.pydata.org/
Matplotlib	An extensive plotting library that enables easy visualization of any form of data analysis. Website: https://matplotlib.org/
Scikit-learn	This library contains all of the project's machine learning algorithms, testing methods, and model selection methods we need. Website: http://scikit-learn.org/stable/index.html

5.2 Identification of parameters of the considered attacks

As a result of the increasing use of web applications by businesses, these companies rely on protecting their sensitive data (such as passwords, bank card numbers, etc.), using firewalls as the first line of protection, but even this is not sufficient to ensure a reasonable level of security and does not guarantee a high level of safety. For this purpose, the second line of protection comes about using vulnerability scanners for intrusion detection systems (IDS) and web applications to ensure that sensitive data is protected from malicious attacks. Systems for intrusion detection may be divided into two main categories:

1. Signature-based intrusion detection systems;

2. Anomaly-based intrusion detection systems.

However, they suffer from many of the shortcomings that make conventional intrusion detection systems ineffective or specific to web applications to detect new attacks. Recently, therefore, many researchers have used methods and algorithms for data mining, machine learning, and artificial intelligence to use their advantages to boost the efficiency of intrusion detection systems, calculated by detection speed and false alarm frequency [11]. The threat can be caused by internal, external, or both external and internal objects. The huge amounts of data stored in the database make it a critical safeguard point for any business-and a valuable object for electronic networks. External threats or intruders act from outside the company and to gain access to your database must overcome your external protection. External threats are restricted by what access they can gain beyond the data network of your organization. Before they can enter the network and access the data available to unprivileged accounts, they must bypass or disable external protection with success. Internal threats or saboteurs work within the company, and can, therefore, bypass external protection. We also have much greater exposure than any external threat, as trusted members of the organization. Another key issue lies in the intent behind the database threat. External threats are almost always malicious: data theft, malfunctions in operation are all potential targets. Internal threats can be equally vicious, including blackmail or

other illicit actions. Within risks, however, are not necessarily malicious. The danger is often not an individual at all, but lax internal policies and security measures that cause accidental or unintended breaches of databases or open up vulnerabilities.

In Section 4, common types of attacks in web applications, such as DDoS attacks, SQL injection attacks, cross-site scripting (XSS) attacks, Zero-day Attacks, Business Logic Attacks, Man-in-the-middle attacks, Malware, Defacements. The two main known vulnerabilities in a web application are SQL injection and cross-site scripting (XSS), which is also the main Internet threat of DDoS attacks, we consider in our research work and analyze them using data sets of these vulnerabilities.

The Database used and considered in this work was obtained at the following links [7], [6] and processing and supplemented by the following dates that are shown in Appendix 1.

DDoS attacks - the types of external threats that most often occur are aimed at suppressing the target web application of the website/server with false traffic, reducing network bandwidth and making it inaccessible to legitimate users.

Cross-site scripting attacks (XSS), also related to external threats, are aimed at users of vulnerable web applications/websites in order to access and manage browsers.

SQL injection attacks are types of internal threats in which an attacker injects malicious SQL code in the form of queries or queries into user input fields in web applications, such as sending forms, contact forms, etc. Thus, they gain access to the internal database of the application, where they penetrate to extract confidential information about customers or the company itself, receive unauthorized administrative access, modify or delete data, etc. or even get full control of the web application.

Disclosure of the parameters of the considered attacks for further analysis.

DDoS attack parameters:

1. Timestamp - request time. Timestamp information is converted to GMT for mobility. To calculate local time, each time stamp must be adjusted;

2. ClientID - a unique integer identifier was set for each client, and due to some privacy issues, these mappings were not issued;

3. ObjectID - a unique integer identifier for the requested URL;

4. Size - number of bytes in the response;

5. Method - the method contained in the client request;

6. Status - this field contains two pieces of information; 2 higher order bits contain the HTTP version specified in the Client request, and the remaining 6 bits indicate the response status code;

7. Type - type of requested file;

8. Server - gives information about which server processed the request.

Parameters of Cross Site Scripting (XSS) Attack and SQL Injection:

1. Length - the first user function to be implemented will be the length of the input;

2. Non-printable characters - non-printable characters are, for example, tabs, line breaks and null characters, that is, characters that are not a written character;

3. Punctuation characters - this function describes the number of punctuation marks in the payload. This includes characters such as “, “>, “>, “/>, which are commonly used in both SQL injection and cross-site scripting attacks;

4. Minimum byte - function that describes the minimum byte in the input. Average minimum byte and standard deviation;

5. Maximum byte - like the minimum byte, creates the function of the maximum byte;

6. Mean byte - a function that describes the average byte of the values of the input characters;

7. Standard deviation byte - a function that describes the average byte of the values of the input characters;

8. Distinct bytes - a function that describes the number of different bytes in the input string;

9. SQL keywords - description of the number of SQL keywords inside the input;

10. Javascript keywords - a function that describes the number of JavaScript-related words for a given input.

The following section will explain how the theory and algorithms in the technical background were used; from finding and extracting good data, to classifying it using Python.

5.3 Data analysis and Model selection

5.3.1 DDoS attacks

According to the DDoS attacks we received, we carried out the following analysis, where we identified the necessary parameters with data cleansing, and applied classifier models that gave certain results.

Using the following code, we derive our data on DDoS attacks, shown in 5.2.

```
np.set_printoptions(precision=3)
pd.set_option('display.float_format', lambda x: '%.3f' % x)
warnings.filterwarnings('ignore')
np.random.seed(8)
%matplotlib inline


def timeit(method):
    def timed(*args, **kw):
        ts = time.time()
        result = method(*args, **kw)
        te = time.time()
        if 'log_time' in kw:
            name = kw.get('log_name', method.__name__.upper())
            kw['log_time'][name] = int((te - ts) * 1000)
        else:
            print('%r  %2.2f ms' % \
                  (method.__name__, (te - ts) * 1000))
        return result
    return timed


data = pd.read_csv('/DDoS_2019_update_dataset.csv')
data_ = data[(data[' Label']!='BENIGN')&(data[' Label']!='WebDDoS')]
len(data_[' Label'].value_counts())
```

Figure 5.2: Data output

The next step will be the identification of the parameters shown in Figure 5.3

	Flow Duration	Total Fwd Packets	Total Backward Packets	Total Length of Fwd Packets	Total Length of Bwd Packets	Fwd Packet Length Max	Pa Le
Flow Duration	1.000	0.003	0.303	-0.045	0.006	-0.152	-
Total Fwd Packets	0.003	1.000	0.096	0.065	0.000	-0.011	
Total Backward Packets	0.303	0.096	1.000	-0.042	0.514	-0.139	
Total Length of Fwd Packets	-0.045	0.065	-0.042	1.000	-0.002	0.015	
Total Length of Bwd Packets	0.006	0.000	0.514	-0.002	1.000	-0.000	
Fwd Packet Length Max	-0.152	-0.011	-0.139	0.015	-0.000	1.000	-
Fwd Packet Length Std	-0.042	0.004	0.014	0.058	0.103	-0.122	
Bwd Packet Length Min	-0.001	-0.000	0.039	-0.001	0.216	0.013	-
Bwd Packet Length Mean	0.002	-0.000	0.210	-0.002	0.676	0.008	
Flow IAT Min	-0.042	-0.006	-0.030	-0.069	-0.003	0.019	-
Bwd IAT Total	0.303	0.006	0.365	-0.011	0.037	-0.039	-
Bwd IAT Min	0.078	-0.001	0.228	-0.017	0.012	-0.057	-

Figure 5.3: Identified Parameters

After creating our data set, cleaning it, and creating functions, the next step was to select and train our models. The selected classifiers are listed below, and they have been trained in all function spaces separately:

1. Logistic regression classifier;
2. Support vector machine;
3. Multinomial Naive Bayes;
4. Random forest;
5. AdaBoost;
6. Artifical neural network;
7. Bag-of-words;

8. Principle component analysis;

9. Decision Trees.

Description of the model with well-trained classifiers and with the exact results that were given during training:

Fit a SVM model to the data, make predictions and summarize the fit of the model. The result with an accuracy of 9.84 percent is shown in Figure 5.4.

Figure 5.4: SVM prediction

The next model is the Logistic Regression, shown in Figure 5.5., which gave the result with accuracy.

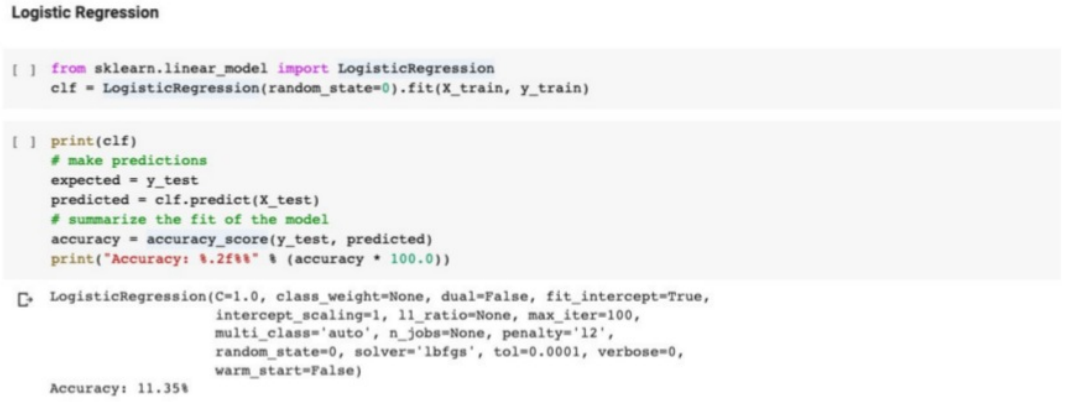

Figure 5.5: Logistic Regression

Decision Tree gave us the result with an accuracy of 72.59 percent, already with an improved prediction than the two models discussed above.The result is visible in Figure 5.6:

Decision Tree

```
from sklearn.tree import DecisionTreeClassifier as dt
clf = dt()
```

```
scores = cross_val_score(clf, X_train, y_train, cv=5, scoring='f1_macro')
```

```
scores.mean()
```

```
0.1608026408303796
```

```
model = dt()
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
print("Accuracy: %.2f%%" % (accuracy * 100.0))
```

```
Accuracy: 72.59%
```

Figure 5.6: Decision Tree

The following model gives a better result with an accuracy of 72.96 percent in our set of DDoS attacks than the above models. The result is shown in Figure 5.7.

```
model = XGBClassifier()
model.fit(X_train, y_train)
```

```
XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1,
              colsample_bynode=1, colsample_bytree=1, gamma=0,
              learning_rate=0.1, max_delta_step=0, max_depth=3,
              min_child_weight=1, missing=None, n_estimators=100, n_jobs=1,
              nthread=None, objective='multi:softprob', random_state=0,
              reg_alpha=0, reg_lambda=1, scale_pos_weight=1, seed=None,
              silent=None, subsample=1, verbosity=1)
```

```
y_pred = model.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
print("Accuracy: %.2f%%" % (accuracy * 100.0))
```

```
Accuracy: 72.96%
```

Figure 5.7: XGBClassifier

5.3.2 SQL Injection and XXS attacks

The first stage is data collection and data processing, hence, the result of the output data is shown in Figures 5.8 and 5.9 with SQL Injection and XXS-attacks.

First 5 lines of SQL

	payload	is_malicious	injection_type
0	'\n	1	SQL
1	a' or 1=1-- \n	1	SQL
2	"a"" or 1=1--"\n	1	SQL
3	or a = a\n	1	SQL
4	a' or 'a' = 'a\n	1	SQL

Figure 5.8: Data of SQL Injection

First 5 lines of XSS

	payload	is_malicious	injection_type
0	data:text/html;alert(1)/*,<svg%20onload=eval(...	1	XSS
1	">*/--></title></style></textarea></script%0A...	1	XSS
2	" onclick=alert(1)//<button ' onclick=alert(1)...	1	XSS
3	';alert(String.fromCharCode(88,83,83))//';aler...	1	XSS
4	">><marquee><img src=x onerror=confirm(1)></ma...	1	XSS

Figure 5.9: Data of XSS attacks

The next step is to determine the parameters. A plot graph of feature importances for better visualization shown in Figure 5.10.

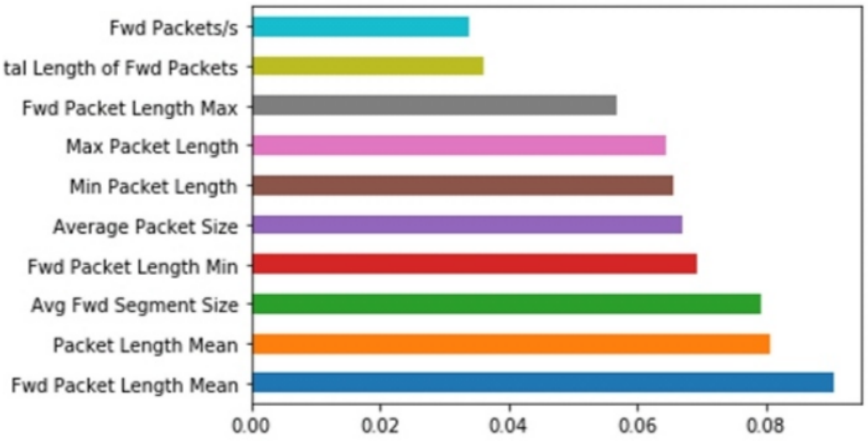

Figure 5.10: Graph of feature importances

Two types of techniques were used for converting the input data to numeric features. Custom features such as length and byte distribution of the data were used as a bag-of-words technique.

Bag of words feature space

We used the N-grams approach to transform our payload data samples into "words" of size N. Example of how the string "<script>" would be transformed when using 1-gram, 2-gram and 3-gram.

N-gram	Payload input	Payload Output
1-gram	"<script>"	['<','s','c','r','i','p','t','>']
2-gram	"<script>"	['<s','sc','cr','ri','ip','pt','t>']
3-gram	"<script>"	['<sc','scr','cri','rip','ipt','pt>']

Figure 5.11: Bag of words feature space

Custom feature space

The good characteristics were selected using the chi-square selection prior to classifier training.

For calculation, we used the formula for the area Chi $\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$.

A Chi-square test is used to check the independence of two occurrences. We can get observed count O and predicted count E given the data of two variables. Chi-Square measures how each other deviates from expected count E and observed count O.

The next move was to pick and train our models after building our data collection, cleaning it out, and defining functions. The classifiers selected are listed above. Most models have good results, with an average overall F1 of 0.9956. The most efficient classifier for all feature spaces was the random forest classifier with an average F1 0.9989 value, as shown in Figure 5.12.

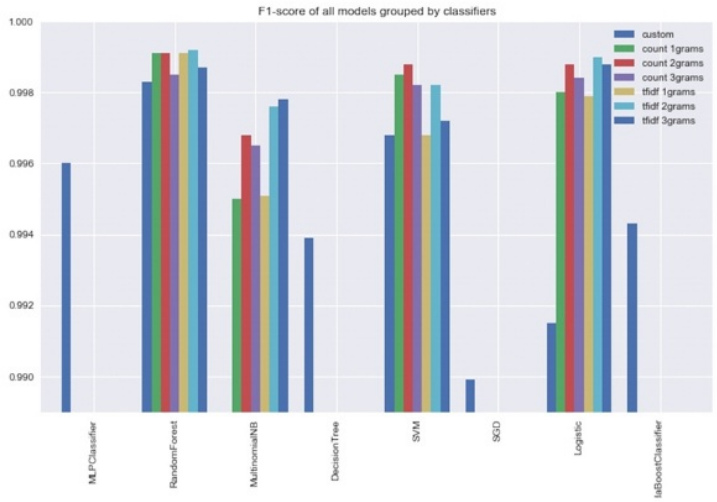

Figure 5.12: All models classifier

In this section, types of threats, identification methods, methods of analysis and forecasting were identified and classified. This allowed us to define the main

criteria for a comprehensive data analysis in the DataSet, which provides modern security solutions for enterprise applications and network systems.

6. Conclusion

From the analysis of existing works presented in the list of literature sources, the absence of a unified information security policy is determined. For example, banks have a separate security policy, universities have a different security policy, there is no general regulation.

Today, machine learning algorithms are used to detect attacks, predict, and prevent attacks. There are solutions using machine learning algorithms Web Application Firewall.

As a result of the analysis of modern security solutions for corporate applications and network systems, the types of threats, identification methods, analysis, and forecasting methods were identified and classified, which allowed determining the main criteria for a comprehensive data analysis in the DataSet.

This paper presents the analysis results for three types of attacks, such as DDoS attacks, SQL-injection, XSS attacks, which pose a great threat to corporate systems and networks, depending on the infrastructure and software applications.

The processes of applying machine learning algorithms in WAF are analyzed, which makes it possible to provide real-time protection for processes and sessions running in a web browser.

Also analyzed DataSets [7], [6], their structures and data types.

Formation of a data model, data cleaning (preprocessing), and determination of parameters for training algorithms, as well as the implementation of algorithms in Python, with the conclusion and description of the results for comparative analysis using machine learning algorithms.

Depending on the organization's infrastructure, its security requirements are defined. The analysis in this dissertation gives the opportunity to formulate new rules, security policies in the organization or adjust them, which is a security model.

References

[1] Hesham Abusaimeh and Mohammad Shkoukani. "Survey of web application and internet security threats". In: *International journal of computer science and network security* 12.12 (2012), pp. 67–76.

[2] LukaS afonov. *Web Application Firewall.* https://habr.com/. May 2009.

[3] Coud Sercurity Alliance. "Top threats to cloud computing v1. 0". In: *White Paper* 23 (2010).

[4] Anthony Bisong, M Rahman, et al. "An overview of the security concerns in enterprise cloud computing". In: *arXiv preprint arXiv:1101.5613* (2011).

[5] cloudflare. *What is SQL injection (SQi)?* https://www.cloudflare.com/. May 2020.

[6] Matt Conran. *Data Sets.* https://vizsec.org/data/. May 2019.

[7] Matt Conran. *DDoS Evaluation Dataset (CICDDoS2019).* https://www.unb.ca/cic/datasets/ddos-2019.html. May 2019.

[8] Matt Conran. *The WAF backed by artificial intelligence.* https://www.networkworld.com/. May 2018.

[9] Matt Conran. *Безопасность предпринимательской деятельности.* https://http://www.aup.ru/books/m6/8_4.htm. May 2009.

[10] John D'Arcy and Pei-Lee Teh. "Predicting employee information security policy compliance on a daily basis: the interplay of security-related stress, emotions, and neutralization". In: *Information & Management* 56.7 (2019), p. 103151.

[11] Nadya ElBachir El Moussaid and Ahmed Toumanari. "Web Application Attacks Detection: A Survey and Classification". In: *International Journal of Computer Applications* 103.12 (2014).

[12] Tristan Fletcher. "Support vector machines explained". In: *Tutorial paper* (2009), p. 4.

[13] Armando Fox et al. "Above the clouds: A berkeley view of cloud computing". In: *Dept. Electrical Eng. and Comput. Sciences, University of California, Berkeley, Rep. UCB/EECS* 28.13 (2009), p. 2009.

[14] Sandro Gerić and Željko Hutinski. "Information system security threats classifications". In: *Journal of Information and organizational sciences* 31.1 (2007), pp. 51–61.

[15] Kathrin Grosse et al. "On the (statistical) detection of adversarial examples". In: *arXiv preprint arXiv:1702.06280* (2017).

[16] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[17] Simon Haykin. *Neural networks: a comprehensive foundation*. Prentice Hall PTR, 1994.

[18] John D Howard. *An analysis of security incidents on the Internet 1989-1995*. Tech. rep. Carnegie-Mellon Univ Pittsburgh PA, 1997.

[19] Jen-Wei Huang, Chia-Wen Chiang, and Jia-Wei Chang. "Email security level classification of imbalanced data using artificial neural network: The real case in a world-leading enterprise". In: *Engineering Applications of Artificial Intelligence* 75 (2018), pp. 11–21.

[20] Martin Husák et al. "Survey of attack projection, prediction, and forecasting in cyber security". In: *IEEE Communications Surveys & Tutorials* 21.1 (2018), pp. 640–660.

[21] ISO/IEC. *ISO/IEC 27001 (ГОСТ Р ИСО/МЭК 27001)*. https://rusregister.ru/. May 2019.

[22] ISO/IEC. *Облачные вычисления)*. https://ru.wikipedia.org/. May 2020.

[23] Erik Johansson and Pontus Johnson. "Assessment of enterprise information security-the importance of prioritization". In: *Ninth IEEE International EDOC Enterprise Computing Conference (EDOC'05)*. IEEE. 2005, pp. 207–218.

[24] Mouna Jouini, Latifa Ben Arfa Rabai, and Anis Ben Aissa. "Classification of Security Threats in Information Systems." In: *ANT/SEIT* 32 (2014), pp. 489–496.

[25] Rafał Leszczyna. "Cost assessment of computer security activities". In: *Computer Fraud & Security* 2013.7 (2013), pp. 11–16.

[26] Ulf Lindqvist and Erland Jonsson. "How to systematically classify computer security intrusions". In: *Proceedings. 1997 IEEE Symposium on Security and Privacy (Cat. No. 97CB36097)*. IEEE. 1997, pp. 154–163.

[27] Katarzyna Łukasiewicz and Sara Cygańska. "Security-oriented agile approach with AgileSafe and OWASP ASVS". In: *2019 Federated Conference on Computer Science and Information Systems (FedCSIS)*. IEEE. 2019, pp. 875–878.

[28] Attacking Techniques Become More. "10 M ajor Security Threats". In: (2009).

[29] Sarah Perez. "The Cloud Isn't Safe". In: *Retrieved June* 3 (2009), p. 2010.

[30] Simon Rogers and Mark Girolami. *A first course in machine learning*. CRC Press, 2016.

[31] Amal Saha and Sugata Sanyal. "Application layer intrusion detection with combination of explicit-rule-based and machine learning algorithms and deployment in cyber-defence program". In: *arXiv preprint arXiv:1411.3089* (2014).

[32] Robert E Schapire. "Explaining adaboost". In: *Empirical inference*. Springer, 2013, pp. 37–52.

[33] Daniel Schatz and Rabih Bashroush. "Security predictions—A way to reduce uncertainty". In: *Journal of information security and applications* 45 (2019), pp. 107–116.

[34] TR Stacey, RE Helsley, and JV Baston. "Identifying information security threats". In: *Information Systems Security* 5.3 (1996), pp. 50–59.

[35] Agnes Talalaev. *8 Types of Cyberattacks a WAF is Designed to Stop*. `https://www.indusface.com/`. May 2020.

[36] Agnes Talalaev. *What is Web Application Firewall (WAF)?)* https://www.webarxsecurity.com/. May 2020.

[37] J Tang et al. "A scalable architecture for classifying network security threats". In: *Science and Technology on Information System Security Laboratory* (2012).

[38] A Tekerek and OF Bay. "Design and implementation of an artificial intelligence-based web application firewall model". In: *Neural Network World* 29.4 (2019), pp. 189–206.

[39] Jinyu Wu, Lihua Yin, and Yunchuan Guo. "Cyber attacks prediction model based on Bayesian network". In: *2012 IEEE 18th International Conference on Parallel and Distributed Systems*. IEEE. 2012, pp. 730–731.

[40] Alice Zheng and Amanda Casari. *Feature engineering for machine learning: principles and techniques for data scientists*. " O'Reilly Media, Inc.", 2018.

[41] Эльдар Бейбутов. *Коробочная безопасность веб-приложений. Внутренност Web Application Firewall*. https://www.anti-malware.ru/. May 2015.

thesis

ORIGINALITY REPORT

24%	17%	12%	16%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.indusface.com Internet Source	2%
2	www.ittoday.info Internet Source	1%
3	Submitted to University of Stellenbosch, South Africa Student Paper	1%
4	slidelegend.com Internet Source	1%
5	nnw.cz Internet Source	1%
6	paper.ijcsns.org Internet Source	1%
7	Submitted to University Of Tasmania Student Paper	1%
8	Submitted to Ajman University of Science and Technology Student Paper	1%

9	arxiv.org Internet Source	1%
10	ejournal11.com Internet Source	1%
11	Submitted to Georgetown University Student Paper	1%
12	A. Suárez Sánchez, P.J. García Nieto, P. Riesgo Fernández, J.J. del Coz Díaz, F.J. Iglesias-Rodríguez. "Application of an SVM-based regression model to the air quality study at local scale in the Avilés urban area (Spain)", Mathematical and Computer Modelling, 2011 Publication	1%
13	en.wikipedia.org Internet Source	<1%
14	www.ijert.org Internet Source	<1%
15	Submitted to Indian Institute of Management, Nagpur Student Paper	<1%
16	Submitted to Middlesex University Student Paper	<1%
17	Submitted to King Fahd University for Petroleum and Minerals Student Paper	<1%

18	Submitted to University College London Student Paper	<1%
19	Submitted to Waterford Institute of Technology Student Paper	<1%
20	Submitted to East Tennessee State University Student Paper	<1%
21	core.ac.uk Internet Source	<1%
22	Timur R. Bikbulatov, Ilya I. Kurochkin. "Simulation of DDoS attack on software defined networks", AIP Publishing, 2019 Publication	<1%
23	Martin Husak, Jana Komarkova, Elias Bou-Harb, Pavel Celeda. "Survey of Attack Projection, Prediction, and Forecasting in Cyber Security", IEEE Communications Surveys & Tutorials, 2019 Publication	<1%
24	Submitted to Aristotle University of Thessaloniki Student Paper	<1%
25	John D'Arcy, Pei-Lee Teh. "Predicting employee information security policy compliance on a daily basis: The interplay of security-related stress, emotions, and neutralization", Information & Management, 2019 Publication	<1%

26	Submitted to Suleyman Demirel University, Kazakhstan Student Paper	<1%
27	Submitted to Hanoi National University Student Paper	<1%
28	Submitted to Melbourne Institute of Technology Student Paper	<1%
29	Jen-Wei Huang, Chia-Wen Chiang, Jia-Wei Chang. "Email security level classification of imbalanced data using artificial neural network: The real case in a world-leading enterprise", Engineering Applications of Artificial Intelligence, 2018 Publication	<1%
30	Submitted to Laureate Higher Education Group Student Paper	<1%
31	Submitted to National University of Singapore Student Paper	<1%
32	export.arxiv.org Internet Source	<1%
33	Adem Tekerek, Omer Faruk Bay. "DESIGN AND IMPLEMENTATION OF AN ARTIFICIAL INTELLIGENCE-BASED WEB APPLICATION FIREWALL MODEL", Neural Network World, 2019 Publication	<1%

34	www.scribd.com Internet Source	<1%
35	Submitted to UT, Dallas Student Paper	<1%
36	Zhewei Wei, Ge Luo, Ke Yi, Xiaoyong Du, Ji-Rong Wen. "Persistent Data Sketching", Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data - SIGMOD '15, 2015 Publication	<1%
37	www.learningace.com Internet Source	<1%
38	Submitted to Astana IT University Student Paper	<1%
39	www.researchgate.net Internet Source	<1%
40	Submitted to University of South Africa Student Paper	<1%
41	"Security, Privacy and Trust in Cloud Systems", Springer Nature, 2014 Publication	<1%
42	Submitted to Swinburne University of Technology Student Paper	<1%

43	Submitted to American Public University System Student Paper	<1%
44	Submitted to Georgia Institute of Technology Main Campus Student Paper	<1%
45	www.cloudflare.com Internet Source	<1%
46	espace.curtin.edu.au Internet Source	<1%
47	old.ieeecss.org Internet Source	<1%
48	Submitted to University of Maryland, University College Student Paper	<1%
49	Submitted to University of Oxford Student Paper	<1%
50	Submitted to General Sir John Kotelawala Defence University Student Paper	<1%
51	www.ijarcce.com Internet Source	<1%
52	Selim Hemissi, Magdy Shayboub A. Mahmoud. "A robust soft margin SVM classifier based on spectral/spatial fusion", Proceedings of the International Conference on Intelligent	<1%

	Information Processing, Security and Advanced Communication - IPAC '15, 2015 Publication	
53	avinetworks.com Internet Source	<1%
54	Submitted to City University Student Paper	<1%
55	Submitted to University of Denver Student Paper	<1%
56	Submitted to Middle East College of Information Technology Student Paper	<1%
57	Submitted to University of Essex Student Paper	<1%
58	epdf.pub Internet Source	<1%
59	www.intechopen.com Internet Source	<1%
60	Submitted to UNITEC Institute of Technology Student Paper	<1%
61	docplayer.net Internet Source	<1%
62	Submitted to Colorado Technical University Online Student Paper	<1%

63	Mustafa I. Jaber, Eli Saber. "Probabilistic approach for extracting regions of interest in digital images", Journal of Electronic Imaging, 2010 Publication	<1%
64	www.ijrte.org Internet Source	<1%
65	dspace.cc.tut.fi Internet Source	<1%
66	Submitted to Prince Sultan University Student Paper	<1%
67	Submitted to University of Liverpool Student Paper	<1%
68	Submitted to Royal Holloway and Bedford New College Student Paper	<1%
69	Submitted to University of Edinburgh Student Paper	<1%
70	Submitted to Metropolia Ammattikorkeakoulu Student Paper	<1%
71	Submitted to University of Hong Kong Student Paper	<1%
72	www.cs.sjsu.edu Internet Source	<1%

73	shura.shu.ac.uk Internet Source	<1%
74	www.ijedr.org Internet Source	<1%
75	www.jcomputers.us Internet Source	<1%
76	hdl.handle.net Internet Source	<1%
77	Submitted to Higher Education Commission Pakistan Student Paper	<1%
78	"Applied Multivariate Analysis", Springer Science and Business Media LLC, 2004 Publication	<1%
79	"Knowledge Discovery for Business Information Systems", Springer Science and Business Media LLC, 2002 Publication	<1%
80	journals.plos.org Internet Source	<1%
81	"Secure IT Systems", Springer Science and Business Media LLC, 2012 Publication	<1%
82	"Global Security, Safety and Sustainability: Tomorrow's Challenges of Cyber Security",	<1%

	Springer Science and Business Media LLC, 2015 Publication	
83	Submitted to 54339 Student Paper	<1%
84	Franck Salles, Reinaldo Dos Santos, Saskia Keskpaik. "When didactics meet data science: process data analysis in large-scale mathematics assessment in France", Large-scale Assessments in Education, 2020 Publication	<1%
85	Submitted to Imperial College of Science, Technology and Medicine Student Paper	<1%
86	Submitted to Coventry University Student Paper	<1%
87	Submitted to iGroup Student Paper	<1%
88	escholarship.org Internet Source	<1%
89	Submitted to The International School Bangalore Student Paper	<1%
90	Submitted to Blackpool and The Fylde College, Lancashire Student Paper	<1%

91	Submitted to University of Bradford Student Paper	<1%
92	Trevor Hastie, Robert Tibshirani, Jerome Friedman. "The Elements of Statistical Learning", Springer Science and Business Media LLC, 2009 Publication	<1%
93	Submitted to Visvesvaraya Technological University Student Paper	<1%
94	Ruijun Cao, Zhen Leng, Shu-Chien Hsu, Wing-Tat Hung. "Modelling of the pavement acoustic longevity in Hong Kong through machine learning techniques", Transportation Research Part D: Transport and Environment, 2020 Publication	<1%
95	Mouna Jouini, Latifa Ben Arfa Rabai. "chapter 16 Threats Classification", IGI Global, 2016 Publication	<1%
96	Martin Husak, Jana Komarkova, Elias Bou-Harb, Pavel Celeda. "Survey of Attack Projection, Prediction, and Forecasting in Cyber Security", IEEE Communications Surveys & Tutorials, 2018 Publication	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		

thesis

GRADEMARK REPORT

FINAL GRADE

/0

GENERAL COMMENTS

Instructor

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

PAGE 11

PAGE 12

PAGE 13

PAGE 14

PAGE 15

PAGE 16

PAGE 17

PAGE 18

PAGE 19

PAGE 20

PAGE 21

PAGE 22

PAGE 23

PAGE 24

PAGE 25

PAGE 26

PAGE 27

PAGE 28

PAGE 29

PAGE 30

PAGE 31

PAGE 32

PAGE 33

PAGE 34

PAGE 35

PAGE 36

PAGE 37

PAGE 38

PAGE 39

PAGE 40

PAGE 41

PAGE 42

PAGE 43

PAGE 44

PAGE 45

PAGE 46

PAGE 47

PAGE 48

PAGE 49

PAGE 50

PAGE 51

PAGE 52

PAGE 53

PAGE 54

PAGE 55

PAGE 56

PAGE 57

PAGE 58

PAGE 59

PAGE 60

PAGE 61

PAGE 62

PAGE 63

PAGE 64

PAGE 65