

	SAKARYA UNIVERSITY JOURNAL OF SCIENCE				 SAKARYA UNIVERSITY
	e-ISSN: 2147-835X http://www.saujs.sakarya.edu.tr				
	Received	Revised	Accepted	DOI	
	14.12.2017	02.07.2018	15.01.2018	10.16984/saufenbilder.365931	

Anomaly Detection Using Data Mining Methods in IT Systems: A Decision Support Application

Ferdî Sönmez ^{*1}, Metin Zontul², Oğuz Kaynar³, Hayati Tutar⁴

ABSTRACT

Although there are various studies on anomaly detection, effective and simple anomaly detection approaches are necessary as the inadequacy of appropriate ways for substantial network environments. In the existing analysis methods, it is seen that the methods of preliminary analysis are generally used, the extrapolations and probabilities are not taken into account and the unsupervised neural network (NN) methods are not used enough. As an alternative, the use of the Self-Organizing Maps has been preferred in the study. In other studies, analysis of data obtained from network traffic is analyzed, here, analysis of other information systems data and suggestions for alternative solutions are given, too. In addition, in-memory database systems have been used in practice in order to enable faster processing in analysis studies, due to the large size of data to be analyzed in large-scale network environments. An analysis of the application log data obtained from the management tools in the information systems was carried out. After anomaly detection results obtained and the verification test results are compared, it is found out that anomaly detection process is successful by 96%. The advantage offered for the company and users at IT and security monitoring processes is to eliminate the need for pre-qualification and to reduce the heavy workload. By this way, it is thought that a significant cost item is eliminated. It is also contemplated that the security vulnerabilities and problems associated with unpredictable issues will be detected through practice and thus many attacks and problems will be prevented in advance.

Keywords: Data Analysis, Anomaly Detection, Artificial Neural Networks, Self-Organizing Maps, In-Memory Database System

1. INTRODUCTION

In the Information and Technology era we are in, the value of the data is getting more and more important every day. One of the most fundamental issues that make data extremely important nowadays is that data can be processed quickly and

advanced data analysis can be done. Depending on the data analysis, more meaningful results are produced and the data are used more effectively [1]. Anomaly detection is one of the crucial purposes of data mining studies [2]. Anomaly detection is used especially in the detection of financial frauds, insurance fraud, cyber security attacks, and failures in critical systems [2]. However, anomaly detection can be used

* Corresponding Author

¹ Istanbul Arel University, Computer Engineering Department, Istanbul, ferdisonmez@arel.edu.tr

² Istanbul Aydın University, Software Engineering Department, Istanbul, metinzontul@aydin.edu.tr

³ Cumhuriyet University, Management Information Systems Department, Sivas, okaynar@cumhuriyet.edu.tr

⁴ Project Manager, Calik Holding, Istanbul, hayatitutar@hotmail.com

effectively in different areas, especially in cancer surveillance and military surveillance of hostile activities [6]. There are several techniques developed for anomaly detection [5][6][7]. One of these techniques is "Clustering Based Anomaly Detection with Artificial Neural Networks" [3].

Data analysis studies have recently been used effectively in log data analysis [6][8]. System log data is becoming more and more critical for systems, applications or devices [7]. This is because the log data constitutes both the memory of the systems, applications or devices to which it belongs, as well as the working and running experience [8][9]. Log data are also considered indispensable for early detection and early warning systems since they contain basic reference information for the safety and decision mechanism of their resources [10]. For this reason, comprehensive and rapid analysis of log data is extremely important [9][11]. There are many studies on the log data analysis. When these studies are examined, it is generally seen that network (intranet / extranet / internet) attacks and security issues are concentrated [6][7][11]. Nevertheless, when analysis techniques are examined, it is often seen that pre-defined techniques are used, extrinsic conditions and probabilities are not taken into account and artificial intelligence methods are not used sufficiently [7][8].

- Focusing on the detection of network attacks in anomaly detection studies, and also the other elements in the systems discussed in this study, reveal the following difficulties for anomaly detection and data mining studies:
- Millions of transactions happen every day. Analyzing this tremendous data load demand for remarkably effective techniques [11].
- The data can be quite distorted. The number of bad events is lower than good events. Usual accuracy-based mining techniques can produce eminently consistent sensors that perform estimates of all normal operations, but these sensors do not detect anomalies [8][9].
- Data labels may not be available immediately and always. Interventions identified as anomalies generally gain awareness after experiencing [8][10].

- It is difficult to monitor the behavior of users. Every type of user (good user, business user, aggressor, fraud) often changes their behavior. Locating new patterns or adjusting patterns is substantial like identifying old patterns [9][12].

Taking the above reasons into consideration, in this study, a log data system covering all the constituent components of the system is constructed, systematic preservation of log data and comprehensive analysis are made using artificial intelligence algorithms, and anomalies are determined by evaluating analysis results. The actual data belong to a substantial network are used to judge the correctness and effectiveness.

In this context, data were analyzed by using Self Organizing Maps (SOM) algorithm and then anomaly was detected based on output data. It is also envisaged that the use of in memory database system (IMDBMS) would be beneficial in order to make transactions as fast as possible and this was taken into account in prototype work. For data analysis, application log data from Windows Management Instrumentation (WMI) on Windows servers in information technology (IT) databases maintained in In-Memory Database (IMDB) tables are used. After the anomaly detection study, it was determined that the anomaly detection process was 96% successful when the results obtained and the verification observation results were compared.

The study was organized as follows. First of all, existing studies are included in the field of anomaly detection. Section 3 discusses the significance of anomaly detection and the detection techniques and development processes. Then, Section 4 discusses different tried-and-tested methods and advantages of using NN and especially SOM in the detection of anomalies. In the next chapter, chapter 5, a recent database management system approach, IMDBMS, is presented in order to speed up the anomaly detection work in a large-scale network environment. The proposed technique and approach is presented in Chapter 6. In addition, chapter 6 presents the methodology followed in the developed application. The results of the test work are presented and discussed in Chapter 7.

2. CURRENT STUDIES

In this section, first of all, recent studies in the national literature are given, followed by recent

studies and findings from the international literature.

In a research conducted to figure out the determinants that reduce the performance of web servers, system errors, analysis of log data and web usage mining have been discussed [11]. In the study, the web server access log records are examined and the classification of error data and the detection of erroneous accesses are emphasized. The identified faults are shared with system administrators and web developers to improve the system, improve the design and help to eliminate mistakes. In the study, it is thought that numerical filtering methods are used but artificial intelligence algorithms and techniques are not used enough.

An anomaly detection method has been emphasized by using an information gain based feature selection method and SOM algorithm in the study carried out on the detection of network based attacks by Anil and Remya [5]. KDD99 (International Information Discovery and Data Mining Competition, 1999) data sets were used for feature selection and evaluating the performance of anomaly-based systems and. Here, genetic algorithm (GA) and SOM have been used to develop feature and information extraction from a large data set like KDD99. The GA has helped identify similar groups from the data set by giving the most important attributes that contribute to abnormal behavior of the links and using the SOM similarity metric. In the anomaly detection process, a number of SOM algorithms, each specialized for a separate attack group, are designed and their performance is measured. In the study, only network-based intrusion detection is considered. The other elements that make up the whole system are not mentioned.

In the study conducted by Ünlü on the investigation of the causes of slowdowns in computer networks, a slowdown based anomaly detection for TCP and DNS over the network is considered [12]. In the study, it was tested the passive examination of the problems causing the computers to slow down from outside the computer and the detection of anomalous situations. When the techniques and methods used are examined, it is seen that statistical analysis methods and accepted test techniques are used (Wilcoxon Rank-sum test, Kolmogorov-Smirnov test and Kullback-Leibler distance test). It has also been determined that NN algorithms are not utilized. The study focuses on networked effects,

which are expressed as external factors that cause the computer to slow down. The most prominent feature of the work is the approach of determining the slowdown problems not through computer-based measurements, but with network-interactive measurements from the outside.

In a study on the detection of cyber attacks, attacks on computer networks and techniques developed for detecting these attacks were discussed and an alternative solution was proposed [13]. Conventional methods of attack detection, the weaknesses and weak points of the methods are mentioned. Methodology of analyzing log data with unsupervised NN algorithms was used. Three intrusion detection methods depending on unsupervised learning algorithms have been tested. As a result of the test runs on the KDD99 dataset, the correct positive rate for attack detection is 34% and the detection rate is 98.7%.

In a study on wireless sensor networks (WSN) in a distributed structure, the status of sensors and anomaly detection on the data from the sensors have been discussed [4]. The focus of the study is on detecting hardware and software faults, unusual events and malicious attacks. Detection procedures were performed by analyzing data from sensors using online anomaly detection techniques. As a result of the study, it is stated that the data analysis for the detection of the anomaly made by the Principal Component Analysis (PCA) of the sensor data is more efficient and effective than the other methods, and misleading rate is minimized.

In a trend analysis study of log data, it was emphasized that the data were analyzed temporally and possible anomaly situations were identified [3]. By analyzing daily data of a call center, it is aimed to determine the seasonal trends and to determine the anomaly situations in the call traffic. Statistical techniques were used in the study, NN algorithms were not used.

Detecting and solving attacks and problems originating solely from the Internet or from the network is not considered sufficient for the healthy operation of the systems and the highest level of business continuity [9][10]. On the other hand, it is essential that all the main components that make up the system are working without problems. For example, the CPU, RAM, disk usage of the servers based on IT systems and the operating status of the operating system are also important and necessary to be stated as essential for the healthy operation

of the systems and to keep the business continuity at the highest level [4][13].

Network intrusion detection techniques are sort by type as anomaly detection and misuse detection [14]. Misuse detection techniques generate attack patterns through labeled diversion supervised learning. These techniques are unable to detect new attacks which are never seen in the training data. However, normal usage patterns are created by anomaly detection techniques. Apparent violations are expected to be detected by inspecting their changes from normal patterns.

It appears that only a single feature of anomaly detection and network management is analyzed in the majority of existing studies [15][16]. However, it is stated that the use of many features in some studies will enrich means of network management [17]. In some studies sharing this idea [18][19][20], it is stated that a few minutes of network analysis is performed and successful results are achieved. However, the applicability of this classical approach is evolving into more difficult on account of the increase in transmission speed of large-scale networks. For example, on Gigabit Ethernet networks, the amount of data transmitted per minute or per transaction is in Terabits. For this reason, with the support of large-scale networks management, more agile approaches are needed that can detect network problems faster.

3. ANOMALY DETECTION TECHNIQUE

Anomalies are behavior patterns that do not fit well-defined normal patterns, or are extraordinary data [15]. Anomalies in a simple two-dimensional dataset are shown in Figure 1. Normal data are concentrated in two main regions, N_1 and N_2 . However, O_1 , O_2 and O_3 appear to be distant anomalous points outside the main region [3].

Anomalies can consist of data that can cause harmful activities such as cyber attacks, frauds in banking and financial systems, terrorist activities, or malfunctioning of a system, or that contain various warnings [15][19][20]. When analyzed, it is easy to see that all of them have common features that are out of order. In anomaly detection, interesting situations or situations that are unusual in real life have key features [13][20].

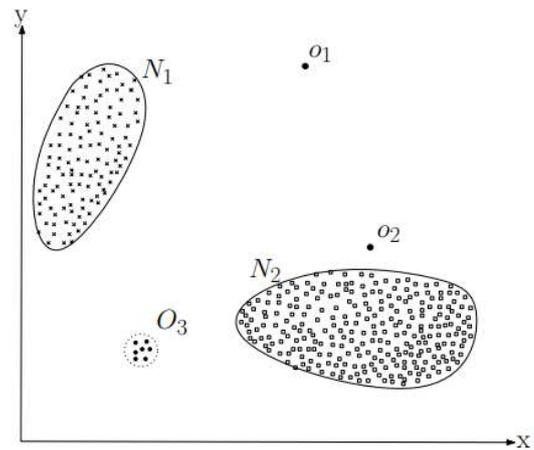


Figure 1. Anomalous Data Example in a Two-Dimensional Data Set [21]

Anomalies can be handled in different types and categories, including point, contextual and collective [21]. Anomaly detection aims to find data on problematic patterns that do not fit the expected behavior [6][19][22]. These problematic patterns often indicate anomalies, discrepancies, incoherent observations, exceptions, deviations, surprises, quirks, or disturbing elements in different application areas. The anomaly detection method has a widespread usage in a wide range of applications like banking, insurance and health service fraud detection, cyber security intrusion detection, critical system failure detection, and various security applications developed for military surveillance of hostile activities [6][15][19].

Identification of anomalies by statistical studies and evaluation of data were first dealt with in the early 19th century [3]. Various research communities have developed different anomaly detection techniques over time [6][13][19][23]. These techniques mostly were developed for specific application areas and infrequently for general application areas [21]. For example, out-of-date credit card transactions may show credit card or identity theft [22]. On the other hand, the presence of abnormally low traffic on a computer network may mean that data having critical care at a certain level are accessed unauthorized on one or more computers and transferred to another location [23].

Generally, each analysis and detection process consists of various process components such as input, analysis and output [24]. There are different process components such as determining the transaction data in accordance with the anomaly type, analyzing the data, labeling the resultant output, and scoring or rating for the sake of achieving the most accurate results in the anomaly

detection process [4]. Different techniques have been developed over time to detect anomalies. For the analyzed data, performing the anomaly detection process by selecting the most appropriate technique according to the anomaly types directly affects the success of the detection process. Commonly used techniques for anomaly detection are as follows [21].

- Statistical Anomaly Detection Techniques
- Nearest Neighbor Based Anomaly Detection Techniques
- Clustering Based Anomaly Detection Techniques
- Information-Theoretic Anomaly Detection Techniques
- Classification Based Anomaly Detection Techniques
- Visualization Based Anomaly Detection Techniques
- Spectral Anomaly Detection Techniques

4. ARTIFICIAL NEURAL NETWORKS AND SOM ALGORITHM

NN, an important component of Soft Computing Techniques, has been developed with a stimulation of information processing manner of the human brain. NN is seen as parallel computation models and consists of parallel and detailed implementation of nonlinear static or dynamic systems [24]. NN stands out with its efficiency in computation and hardware applications. Generalization ability is a significant advantage of NN. Thanks to their ability they can sort new patterns with high accuracy [25]. A major disadvantage in addition to the advantages of the NN is that it is harassed by its interpretations [26][27][28]. If a NN model can be created and trained in a short time, it is also possible to determine the best model with a large number of trials [24][26]. NNs are classified in various forms within themselves. Highlights from these classifications are according to structures, architectural layers, learning algorithms and usage purposes [24].

It has been proposed in many studies that NN is beneficial in non-linear nonvolatile situations and in data structure [29]. This quality seems to be an important reason for the use of NN. In this study, SOM was chosen as the method of NN. The

advantage of this method is that it is used not only as a general fraud and anomaly detection method, but also for certain areas or applications, and can be easily used in every information system in which log records are sequentially recorded [24]. In addition, this approach is an unsupervised technique [30] and therefore avoids the problems associated with inadequate training data affecting the results of anomaly detection of supervised data mining methods [31]. The SOM, which constitutes the application of this work, has been examined in detail under a separate section.

4.1. Self-Organizing Maps

SOM is a NN model developed by Teuvo Kohonen in 1982 that uses a relational memory foundation and learns without supervision [30]. Relational memory is the form of learning that is related to the strength of the bond between two objects [31]. The SOM model takes a data set that has no idea about it in advance and provides clustering according to the correlations and distances between them, and it creates a visual map at the end of the process [30].

SOM Networks consists of four main components: initiation, competition, cooperation, adaptation [32]:

Initialization: All input data are randomly weighted with small values to start the process.

Competition: For each input data, competition is started to calculate its own values by providing a basic decomposition function to the neurons. The winning neuron is identified and declared with the smallest value of the decomposition function.

Cooperation: The spatial location of other neurons is determined according to the winning neuron, and the cooperation between neighboring neurons is established. Thus, winning neurons form topological neighbors with the neurons closest to them.

Adaptation: According to the connection relations between the weights of the input data, the values of nearby neurons are reduced by the decomposition function so that the compatibility of input data similar to the winning neuron is extended. Thus, compatibility is extended by applying input data to the next winning neuron.

4.2. Training of the Self-Organizing Map Networks

SOM is a single-layer network and involves input and output neurons. The number of input neurons is determined by the number of variables in the dataset. Each of the output neurons represents a cluster [32]. The arrangement of neurons in the output layer is also very important. The array may be linear, rectangular, hexagonal or cubic depending on preference. Generally, rectangular and hexagonal arrangements are preferred [31]. However, the rectangular array is usually implemented as a square array. In terms of topological neighborhood, the arrangement at there is important [32]. There is no direct connection between the output neurons. The connection between input neurons and each output neuron is shown by reference vectors [32]. Since SOM networks are a single-layered network, there is no question that the data to be used in education are dependent variables. For this reason, it seems to be ideal for solving problems related to clustering analysis [33]. However, expert opinion is needed to control the accuracy of the results obtained.

4.3. Evaluation of SOM Algorithm and Other NN Algorithms

SOM networks are generally preferred for clustering and visualization of data [33]. SOM networks reduce multidimensional data sets to two-dimensional maps [34]. The reference vectors generated for each cluster come together to form a two-dimensional map. The topological neighborhoods on this map show the relations between the clusters. For this reason, the use of SOM networks has increased in recent years [33]. Many types of SOM networks have been developed [34].

Kangas et al. have proposed several ways to improve dynamic topological neighborhood and reference vectors by assigning initial values to reference vectors, Learning Vector Quantization (LVQ) algorithm [34]. Instead of LVQ in the comparison of unsupervised SOM and supervised LVQ algorithms, Pal et al. introduced a generalized learning vector quantization (GLVQ) algorithm [35]. LVQ is separated from SOM depending on whether it is supervised and does not have topological neighborhood. Only the winning neuron coefficients are updated in the LVQ algorithm, while other neuron coefficients are

updated in the GLVQ algorithm [36]. Martinetz et al. developed a Neural-Gas algorithm that uses the SOM algorithm and the K-Means method as alternatives to the prediction of time series [37].

4.4. Using Self-Organizing Maps in The Anomaly Detection

SOM is often used for clustering. The architecture of the SOM consists of a feed-forward NN consisting of neurons arranged in a single layer and a rectangular array. When an input pattern is presented to the SOM, each neuron calculates how similar it is to the weight of the input. The most weighted neuron (minimum distance in the input field) is advertised as a competition for entry motifs, and the weight of the winning neuron is reinforced to reflect the end result. Winning neurons receive the most learning at any stage; as neighbors receive less, they are moving away from the winning neuron [38].

Here, a kernel similarity criterion is used instead of the distance of the conventional total squares in order to measure the distance. Kernel functions make it possible to clustering complex data as linearly as it is in this study, easily and quickly [39].

A kernel function is adopted for measuring the distance between two data points [39]. When a set of data (X) is given in a D-dimensional real space (R_d), a nonlinear mapping function from the input field into a high dimensional feature space (H) is as follows:

$$\varphi: R^d \rightarrow H, \quad \vec{x}_i \rightarrow \varphi(\vec{x}_i) \quad (1)$$

Here:

$$\vec{x}_i = [x_{i,1}, x_{i,2}, \dots, x_{i,d}]^T \quad (2)$$

and

$$\varphi(\vec{x}_i) = [\varphi_1(\vec{x}_i), \varphi_2(\vec{x}_i), \dots, \varphi_H(\vec{x}_i)]^T \quad (3)$$

By applying the mapping, a point \vec{X}_i^T, \vec{X}_j becomes $(\vec{X}_i) \cdot \varphi(\vec{X}_j)$. The basic idea for kernel-based learning is that there is not any need to certainly specify the φ function [39]. Therefore, the kernel distance between two patterns (\vec{X}_i and \vec{X}_j) can be given as:

$$\begin{aligned} \|\varphi(\vec{x}_i) - \varphi(\vec{x}_j)\|^2 &= (\varphi(\vec{x}_i) - \varphi(\vec{x}_j))^T (\varphi(\vec{x}_i) - \varphi(\vec{x}_j)) \\ &= \varphi^T(\vec{x}_i) - \varphi(\vec{x}_i) - 2\varphi^T(\vec{x}_i) \cdot \varphi(\vec{x}_j) + \varphi^T(\vec{x}_j) \cdot \varphi(\vec{x}_j) \\ &= K(\vec{x}_i, \vec{x}_i) - 2K(\vec{x}_i, \vec{x}_j) + K(\vec{x}_j, \vec{x}_j). \end{aligned} \quad (4)$$

Gaussian kernels (also known as radial-based functions) are known for their better linearity in classification of linear and polynomial kernels in most of the test problems [39]. Gaussian kernels can be expressed as:

$$K(\vec{x}_i, \vec{x}_j) = \exp\left(-\frac{\|\vec{x}_i - \vec{x}_j\|^2}{2\sigma^2}\right), \quad \sigma > 0 \quad (5)$$

For Gaussian kernels it is possible to simplify $K(\vec{x}_i, \vec{x}_j) = 1$ and hence the equation (4) becomes:

$$\|\varphi(\vec{x}_i) - \varphi(\vec{x}_j)\|^2 = 2(1 - K(\vec{x}_i, \vec{x}_j)) \quad (6)$$

This measurement is used in our algorithm.

Finally, when all inputs are assigned to a particle, they have as many as the total number of particles. Therefore, these clusters need to be categorized. The purpose here is to find the anomaly in the input data.

Less-membership clusters are good candidates for 'abnormal' clusters. The averages of the members of all the clusters were used for this categorization. Members have been searched to categorize clusters. However, for this categorization the averages of the members of all the clusters were used. Members who are too far from the average of their clusters are not normal. False positive, false negative, true positive and false positive numbers were used to measure the accuracy of our method [36][40]. In order to measure the performance of the developed algorithm based on these results, Receiver Operating Characteristic (ROC) was used. ROC has been chosen as a traditional and accepted method to investigate the performance of anomaly detection methods [41][42][43].

5. IN MEMORY DATABASE SYSTEMS

The collection, storage and reporting of data throughout history has been of great importance. Along with the emergence and development of computer technologies, Relational Database Systems (RDBMS) have begun to be developed, especially since the 1970s, to consistently capture report and analyze data [44]. RDBMS has been developed everyday and has been up to date and continues to be used effectively. However, due to the increase in data size in recent years, new needs arise that the duration of analysis and reporting is getting longer and new generation applications are being accessed in different media and formats, making the development of new generation

database systems indispensable. In the early years of the 2000s, studies began to develop new generation database systems [44][45]. Along with the spread of the concept of social media, this process has been accelerated [44].

One of the next generation database systems developed considering the above-mentioned reasons is In-Memory Database Systems (IMDB). These database systems, which are briefly referred as "in-memory", have been developed with reference to the basic features of RDBMS. Therefore, IMDBMS has its essential features, which are considered to be indispensable of the IMDB and called ACID (Atomicity - Consistency - Consistency, Isolation, Durability) in its own structures [46][47]. The comparison of IMDB and IMDBMS is shown in Figure 2 in general.

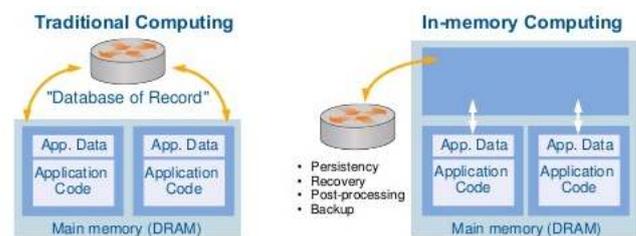


Figure 2. Traditional Database Systems and In-Memory Database Systems

The most basic feature that distinguishes IMDBMS from other approaches is that it stores the data in RAM. Thus, it is suggested that during database operations, disk I / O operations are not needed and that it will speed up the reporting and analysis processes [45][47]. For example, a process performed in seconds can be processed for a period of milliseconds [45][46]. The second fundamental characteristic of IMDBMS is that it adopts column-based data and index structure instead of row-based data and indexes. Thanks to this feature, IMDBMS has become a database system that takes up less space [46][48]. These rates vary according to the products of the companies that develop IMDBMS, the structure of the data, the architectures and technologies used [45][49].

6. APPLICATION

System data in an analyzed dataset can be achieved by visualizing the SOM map using a binary classification algorithm [34][50][51]. Unusual system behaviors seen in SOM are treated as anomaly values in our study. Therefore, the anomaly detection in our study is based on the problem of the detection of the anomaly.

A simple threshold type binary classification technique is used as a useful tool for detecting anomalies. SOM algorithm (in two-dimensional space) emerges as an effective approach and has computational simplicity which constitutes it as a competent method [51]. It is also possible to use different mathematical classification algorithms that do not concentrate on the classification problem in the study. Nevertheless, it is necessary to discern the normal and abnormal behaviors of the system. For this reason, it is aimed that the classification threshold and the possible anomaly behaviors are reflected in the SOM map in the most accurate way. It is therefore aimed to classify operations using the following function, which returns a logical true value when abnormal system operations are performed and returns a logical false value when normal system operations are performed [34]:

$$\varphi\alpha_i = \begin{cases} \text{true for } d(ci, c) > \tau \\ \text{false for } d(ci, c) \leq \tau' \end{cases} \quad (7)$$

Where the α_i represents i th element of the system, i , c_i represents the element of system i in the SOM map, c represents the center of the entire SOM map, d represents a selected degree of uniqueness and s classification threshold.

Here, the threshold type binary classification algorithm was used by way of the efficacy simplicity, and experimentally controlled effect in the calculation [50]. It is, however, quite convenient to obtain ROC curves that simply evaluate the approaches that are being investigated by shifting the threshold value.

In this study, six dimensions (CPU usage rate (%), RAM usage rate (%), incoming network traffic - Mbps / sec, outgoing network traffic - Mbps / sec, fixed by considering the reasons and methodology explained above and in the previous sections, a new network anomaly detection system ILogAnalyzer which aims to detect and inform network and system anomalies based on daily log (log) analysis of disk usage rate -%, system usage rate -% is presented. The main contribution of the presented system can be listed as simultaneous analysis of six dimensions, detection of different anomalies and generation of specific stimuli. It also provides important information that allows the network administrator to identify the root and application (anomaly) of the anomalies detected to aid in the detection of anomalies and the development of countermeasures.

The process [42][43] followed by the anomaly detection infrastructure is summarized in the study steps of the application, briefly as follows:

- A) Selecting the classified data and executing the appropriate changes.
- B) Creating and optimizing the model using SOM and training data.
- C) Implementing the model to inspect the order and information on a regular or irregular basis.

Thus, the SOM model created the maps that hold the dissimilarities in data. An akin approach has been implemented at each node. If an anomaly is detected in this process, consistent cautions will be generated and a way of modeling the corresponding prevention systems will be opened to prevent processes / network operations.

IMDBMS is used to keep log data to be analyzed for reasons mentioned in the section on existing work and in-memory database systems. The anomaly detection process is performed on the analyzed data and the level of detected anomalies is determined and informed by the mapping method on the analysis result screen of the application.

6.1. Data Set

The dataset used in our research contains actual data (log records) obtained from a telecommunication company's network and systems of the main campus in Istanbul for a period of 2017 and is managed by the network management system manager of the firm and by a network specialist and processed to a preliminary and management process till it was passed. Raw data includes approximately 100 megabytes of compressed network dump and system log (log) data from a 7 day network traffic. System or network operations are labeled as normal or anomalous. If the anomaly is a network attack, it is labeled with the attack type specified.

6.2. Architectural Structure and Components of the Application

Log data obtained from the systems were stored in IMDBS, then the resource usage status of the systems was examined with the aid of SOM algorithm for the purpose of anomaly detection. Visualization of the results by graphic and mapping method and production of meaningful results were performed by ILOGAnalyzer.

Detection of anomaly in practice is basically carried out in two stages. First, the test data was used and appropriate input parameters were found, and then the program was applied to the unlabeled data using the appropriate input parameters (the input parameters that delivered the best result). The system uses a training process to derive input parameters from the test data and perceives inputs as normal or abnormal. Designed as a Windows form application, ILogAnalyzer is developed in C# language and .Net platform. As the application database, Microsoft SQL Server 2016 Developer version is used. The actual application log data, belonging to the Windows servers in the IT systems received via WMI, has been transferred to the IMDB tables created for testing purposes. The log data was passed through the necessary examinations (masking etc.) based on the corporate information security and privacy policies before the tabulation was transferred. The number of records in the test tables ranges from 1,000 to 10,000. The database connection to the application is provided using the ADO.NET DB connection method. As the .Net Framework, version 4.5 is preferred. In database table querying, the TSQL query language is used.

6.3. Steps of Application

The steps of the ILOGAnalyzer application are designed as follows;

- a) Processing of the log data set to be analyzed
- b) Normalization of data (normalization)
- c) Starting the analysis process with the SOM algorithm
 - i. Determination of proximity and neighborhood relations between data
 - ii. Clustering and classification of data
 - iii. Clustering and post-categorization tagging
- iv. Completing the mapping of data
- d) Initiation of anomaly detection
 - i. Acquisition of analysis and mapping results data by SOM algorithm
 - ii. Identification of data patterns of clusters
 - iii. Determination of the central position of each cluster
 - iv. Determination of the positions of other data relative to the detected central position

- vi. Detection of anomalies for data which are distant from the center position depending on the distance measurement
- e) Completion of analysis and anomaly process and display of result
- f) Marking of anomalous data by coloring.

6.4. Usage of Application

The steps for using ILOGAnalyzer are summarized as follows;

- a) Analysis of the log data to be displayed on the screen: The source of the log data to be analyzed is determined and the data table or file to be analyzed is selected from this source to load the data sample on the screen. (Figure 3)
- b) Selection of log data to be analyzed: From the log data sample loaded on the screen, the column headings of the data to be included in the analysis process are marked and the data set for analysis is created with the selected log data. (Figure 3)
- c) Configuration of log analysis preferences: Preferences are made for the SOM algorithm which will analyze the log data (Figure-4 and Test Procedures and Assessment of Results section)
- d) Starting the log analysis process: After the data set is created and the analysis preferences are made, the analysis process is started.
- e) Displaying the analysis result: After the analysis process is complete, the anomaly detection result list displayed (Figure 5).
- f) Analysis of anomaly detection map and graph: From the graphs, maps and lists in the analysis result screen, the log analysis process and anomaly detection results can be examined (Figure 3, Figure 4, Figure 5 and Test Procedures and Evaluation of Results section).

The screenshot shows the application's main window with several sections:

- Data Seçim Parametreleri:** Includes 'DB Tablo Seçimi' (VolumeUsage), 'DB Data Yükle' (Kayıt Limiti: 1500), 'CSV Data Dosyası', and 'Text Data Dosyası'.
- Analiz Parametreleri:** Includes 'DB Data Yükle' (ID, NodeID, VolumeID), 'CSV Data Yükle' (Date Time, Disk Size, AvgDiskUsed), and 'Text Data Yükle' (Tümünü Seç).
- Analiz Parametreleri:** Includes 'Analiz Parametreleri' (Analiz Parametreleri) and 'Analiz Parametreleri' (Analiz Parametreleri).
- Orjinal Veri Seti:** A table with columns: ID, NodeID, VolumeID, Date Time, DiskSize, AvgDiskUsed, MinDiskUsed.

ID	NodeID	VolumeID	Date Time	DiskSize	AvgDiskUsed	MinDiskUsed
275083	365	333	21.05.2017	17179336704	12566541994.66...	1165508601
275084	365	334	21.05.2017	18440400896	15161410176	1387028071
275085	365	335	21.05.2017	128740016128	100728204596.0...	1003309099
275086	365	336	21.05.2017	0	-2	-2
265109	365	333	5.06.2017	17179336704	14831213056	142728929
265110	365	334	5.06.2017	21358641152	18243388330.66...	177362452
265111	365	335	5.06.2017	128740016128	109450450200.1...	109255163
265112	365	336	5.06.2017	0	-2	-2
267515	365	333	6.06.2017	17179336704	14708537258.66...	143012085
267516	365	334	6.06.2017	21821640704	18275453781.33...	1787649221
267517	365	335	6.06.2017	128740016128	1110064379687.8...	109568827
267518	365	336	6.06.2017	0	-2	-2

Figure 3. Selection and Preparation of Log Data to be Analyzed in the Application

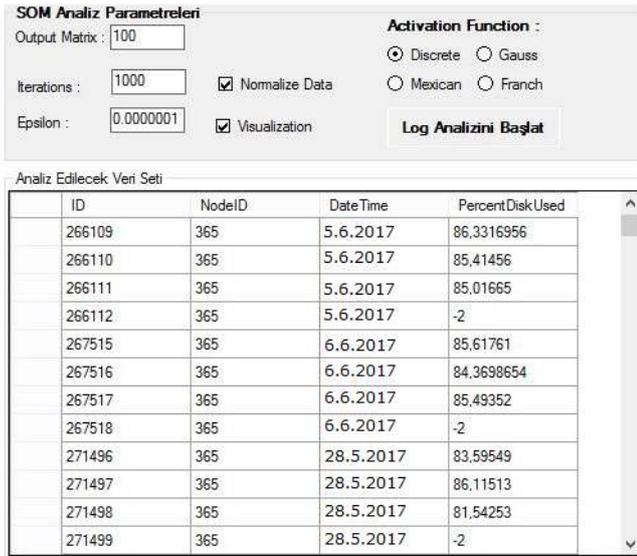


Figure 4. Implementation of Log Analysis Preferences in the Application



Figure 5. List of Data Found as Anomaly in Practice

7. EVALUATION OF THE RESULTS

After the analysis process is completed, the results of analysis are evaluated and the results of the anomaly detection are evaluated by examining the generated analysis graph and SOM map. Experiments were performed with real data set and related anomalies were determined. The aim here is keeping the correct positive rate at a certain level and to detect the highest number of anomalies. Table 1 shows the test results.

ROC was used as the evaluation criterion for the results obtained. ROC has been chosen as a traditional and accepted method of investigating the efficiency of anomaly detection methods [41]. In this case, the ROC curves are better tools for performance comparison. ROC curves display the sensitivity and false positive rate.

$$Precision = \frac{true\ Positives}{true\ Positives + false\ Positives} \quad (8)$$

$$False\ Positive\ Rate = \frac{False\ Positives}{True\ Negatives + False\ Negatives} \quad (9)$$

The point (0, 1), which is regarded as the perfect point in the ROC graph, corresponds to the

perception of all anomalies with zero false alarm rate (equation 12). Depending on this reason, the perfect ROC curve contains this point. This result is accepted as the better detection method as the ROC curve approaches [51]. On the other hand, (1, 0) refers to the lowest detection (Equation 11) performance. The area under ROC curve (AUROC), which is only applied for measuring the area under the curve, is adopted to assess a specific curve [50]. The best value for AUROC is 1. The two important values the assessing the curve are the lowest false alarm rate with correspondence to the lowest anomaly (1) detection rate (LADR) and the highest anomaly detection rate with correspondence to the zero false alarm rate (ZFAR). For ZFAR the ideal value is 1 and for LADR the ideal value is 0 [41].

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Pos. + True\ Neg. + False\ Pos. + False\ Neg.} \quad (10)$$

$$Recall = \frac{true\ Positives}{True\ Positives + False\ Positives} \quad (11)$$

$$False\ Alarm\ Rate = \frac{false\ Positives}{True\ Positives + False\ Positives} \quad (12)$$

As an additional criterion in appraising the efficiency of the recommended approach, F1 Score defined in [52] was used:

$$F1\ Score = 2 \times \frac{precision * recall}{precision + recall} \quad (13)$$

The most effective approach is to diminish the false alarm rate by any means possible and meantime enhance the anomaly detection rate [51][52]. The parameters of the corresponding classification algorithm of the referenced SOM clustering-based method are arranged to generate ROC curves [51][52]. For each of the data sets examined, the Sensitivity, Precision, Accuracy and F1 score of the SOM method were calculated at a specific point in the ROC curve. By using the method explained in [53], the classification threshold was determined (equation 7).

After the analysis performed, the data formed a clustering as shown in Figure 6. The concentration of anomalies was determined in the A1, A2 and A3 regions [54]. Another result is that the SOM algorithm's learning coefficient / iteration residuals, the analysis process and the anomaly detection process are more successful.

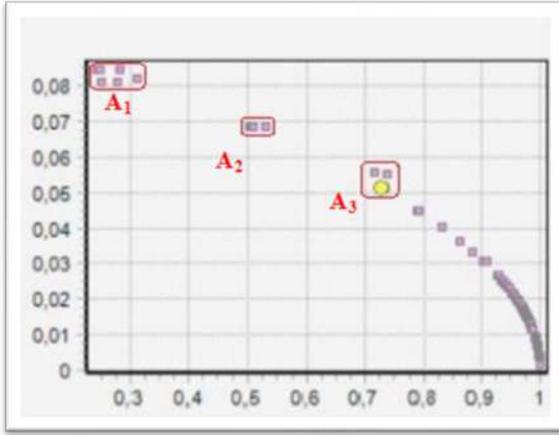


Figure 6. Anomaly Detection Application Sample Graph View

The results of the experiments performed according to the SOM-clustering method and the ROC curves are shown in Figure 6. The following results were obtained (Table 1 and Table 2) following the decision on the classification threshold of the mentioned method for gaining the optimal values for the fixed value and on the classification algorithm of the reference method selected according to the method described in the application section. It has been observed that the approach suggested in the study would return the false alarm ratio close to zero and the true alarm ratio close to 1 [1][40][55], in order to provide acceptable results when considering other similar studies. Sensitivity, Precision, Accuracy, F1 score and AUROC (Figure 7) values are taken into account in the determination of the superiority mentioned here [55][56]. The complexity analysis is displayed in Table 1 and the evaluation results are displayed in Table 2.

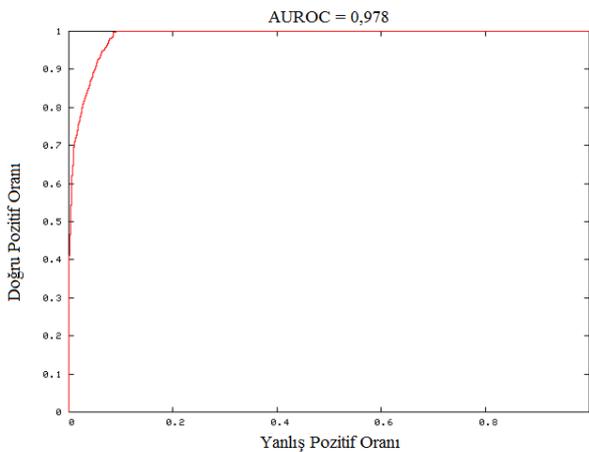


Figure 7. ROC Curve for Anomaly Detection

Table 1. Confusion Matrix

	Positive	Negative
Positive	2483	112
Negative	441	3368

Table 2. Test Results

Parameters	Values
Training Cluster	57641
Test Cluster	6404
True Classified	5851
False Classified	553
True Positive Rate	0,85
True Positive Number	2483
True Negative Number	3368
False Alarm Rate	0,12
False Positive Number	441
False Negative Number	112
Accuracy	0,91
Perception	0,96
Specificity	0,13
F1 Score	0,93
AUROC	0,98

It has been observed that the accuracy of the results obtained from the developed application is close to or higher than the accuracy of many anomaly detection studies that have been successfully demonstrated by using NN or statistical techniques for similar purposes [1][4][5][40][55][56]. It has also been observed to be close to or less than the specificity obtained from successful studies [1][40][42][55][56]. In this case, it can be said that the developed application has achieved an anomaly detection performance at an acceptable level.

8. CONCLUSION AND EVALUATION

In this study, an approach and application for anomaly detection in computer networks and systems is proposed. In practice, SOM technology is used to visualize the operations in the computer network and systems. Anomalies are detected by utilizing the threshold type binary classification algorithm after reflection of behaviors in the observed systems to SOM.

In practice, application logs received from WMI on Windows servers in IT systems that are kept in

IMDB tables are used. First, the data were analyzed using SOM algorithm and then the anomaly was detected based on the output data. After the anomaly detection study, when the results obtained and the results of the verification observations were compared, it was found that the anomaly detection process was 96% successful.

The application is aimed at alleviating the intensive workload on users as it removes the need for pre-definition of IT systems tracking and security tracking. In this way, it may be possible to alleviate the financial burden by eliminating a significant cost for enterprises. On the other hand, anomaly situations, security vulnerabilities and system problems that can be caused by unidentified or unlikely issues can be easily detected with this application, so that many problems that may arise can be recognized beforehand and countermeasures can be put into practice.

Another benefit of the presented method is that it is a generally accepted anomaly detection method, in other words it can be used conveniently for any information system which gathers data obtained from daily log records and not to specific areas or applications. Furthermore, the application developed based on the presented method is an application based on an unsupervised method that avoids problems related to inadequate training data that substantially affect the final determination results of a consultative data mining method.

The most important advantage for users of the application is that they can analyze and distinguish unusual data from the anomaly in the data set by analyzing data that have no idea about it in advance. With this feature, users who are working in the field of IT security and monitoring are offered the opportunity to remove the system, device and applications they want to monitor, from the one that has already been defined to individual monitoring systems. Therefore, it is thought that it is possible to perform security and monitoring operations efficiently and easily in all system, device and application services which are not additionally defined.

Considering its highly efficient and accurate anomaly detection and identification capabilities, the proposed system is thought to represent a new and effective system that assists the management processes of big size IT systems and networks and guarantees the quality and availability of services.

It is aimed to enhance the performance of the proposed ILogAnalyzer in various applications and services by analyzing the behavior of different scenarios and improving the early warning and prevention system for security and monitoring applications on IT systems with forthcoming studies.

REFERENCES

- [1] M. V. O. Assis, J. J. P. C. Rodrigues, M. L. Proença. "A seven-dimensional flow analysis to help autonomous network management", *Information Sciences*, 278, 900-913, 2014, doi: 10.1016/j.ins.2014.03.102.
- [2] A. Coluccia, A. D'Alconzo, F. Ricciato. "Distribution-based anomaly detection via generalized likelihood ratio test: A general Maximum Entropy approach", *Computer Networks*, 57(17), pp.3446-3462, 2013, <http://dx.doi.org/10.1016/j.comnet.2013.07.028>.
- [3] F. Mata, P. Żuraniewski, M. Mandjes, M. Mellia. "Anomaly detection in diurnal data", *Computer Networks*, 60, pp. 187-200, 2014.
- [4] M. A. Rassam, A. Zainal, M. A. Maarof. "An Efficient Distributed Anomaly Detection Model for Wireless Sensor Networks", *AASRI Procedia*, 5, pp. 9-14, 2013, doi: 10.1016/j.aasri.2013.10.052.
- [5] S. Anil, R. Remya. "A hybrid method based on genetic algorithm, self-organised feature map, and support vector machine for better network anomaly detection", *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*, Tiruchengode, pp.1-5, 2013, doi: 10.1109/ICCCNT.2013.6726604
- [6] A. Juvonen, T. Sipola, T. Hämäläinen. "Online anomaly detection using dimensionality reduction techniques for HTTP log analysis", *Computer Networks*, 91(14), pp.46-56, 2015, doi: 10.1016/j.comnet.2015.07.019.
- [7] I. Fronza, A. Sillitti, G. Succi, M. Terho, J. Vlasenko. "Failure prediction based on log files using Random Indexing and Support Vector Machines", *Journal of Systems and*

- Software*, 86(1), pp.2-11, 2013, doi: 10.1016/j.jpp.2012.06.025.
- [8] D. Olszewski. "Fraud detection using self-organizing map visualizing the user profiles", *Knowledge-Based Systems*, 70, 324-334, 2014, doi: 10.1016/j.knosys.2014.07.008.
- [9] C. Modi, D. Patel, B. Borisaniya, H. Patel, A. Patel, M. Rajarajan. "A survey of intrusion detection techniques in Cloud", *Journal of Network and Computer Applications*, 36(1), pp.42-57, 2013, doi: 10.1016/j.jnca.2012.05.003.
- [10] A. Botta, A. Dainotti, A. Pescapé. "A tool for the generation of realistic network workload for emerging networking scenarios", *Computer Networks*, 56(15), pp.3531-3547, 2012, doi: 10.1016/j.comnet.2012.02.019.
- [11] D. A. S. Resul, I. Turkoglu, I., M. Poyraz. "Analyzing of system errors for increasing a web server performance by using web usage mining", *IU-Journal of Electrical & Electronics Engineering*, 7(2), pp.379-386, 2007.
- [12] S. A. Ünlü. "Ağ Üzerinden Yavaşlama Tabanlı Anomali Tespiti", *Tez Çalışması*, TOBB Ekonomi ve Teknoloji Üniversitesi, Fen Bilimleri Enstitüsü, 2011.
- [13] P. Ma. "Log Analysis-Based Intrusion Detection via Unsupervised Learning", *Master of Science*, School of Informatics, University of Edinburgh, 2003.
- [14] C. Chiu, Y. Ku, T. Lie, Y. Chen. "Internet auction fraud detection using social network analysis and classification tree approaches", *International Journal of Electronic Commerce*, 15 (3), pp.123–147, 2011.
- [15] A. Li, L. Gu, K. Xu. "Fast Anomaly Detection for Large Data Centers," *2010 IEEE Global Telecommunications Conference GLOBECOM*, Miami, ABD, 2010, doi: 10.1109/GLOCOM. 2010. 5683551
- [16] Y. Kanda, K. Fukuda, T. Sugawara. "A Flow Analysis for Mining Traffic Anomalies", *2010 IEEE International Conference on Communications, Cape Town*, 2010, doi: 10.1109/ ICC.2010. 5502463
- [17] S. Molnar, Z. Moczar. "Three-Dimensional Characterization of Internet Flows," *2011 IEEE International Conference on Communications (ICC)*, Kyoto, 2011, doi: 10.1109/icc. 2011.5963476
- [18] P. P. Cortez, M. Rio, M. Rocha, P. Sousa. "Internet Traffic Forecasting using Neural Networks," *The 2006 IEEE International Joint Conference on Neural Network Proceedings*, Vancouver, BC, 2006, doi: 10.1109/IJCNN. 2006.247142
- [19] A. Lakhina, K. Papagiannaki, M. Crovella, C. Diot, E.D. Kolaczyk, N. Taft. "Structural analysis of network traffic flow"s, *SIGMETRICS Perform. Eval. Rev.*, pp.32, 61–72, 2004.
- [20] C. Yoohee, K. Yihan. "Case study of an anomalous traffic detection on the aggregation points of enterprise network", *International Conference on Advanced Communication Technology (ICACT)*, Seoul, 2011.
- [21] A. Chandola, V. Chandola, V. Kumar. "Anomaly Detection: A Survey", *ACM Comput. Surv.*, 41(3), 2009, doi: 10.1145/1541880. 1541882
- [22] N. Carneiro, G. Figueira, M. Costa. "A data mining based system for credit-card fraud detection in e-tail", *Decision Support Systems*, 95, pp.91-101, 2017, doi: 10.1016/j.dss.2017.01.002.
- [23] V. Kumar. "Parallel and distributed computing for cybersecurity" *IEEE Distributed Systems Online*, 6(10), 2005. doi: 10.1109/MDSO. 2005.53 .
- [24] S. Haykin, *Neural Networks and Learning Machine*, 3E, *Pearson Education Inc.*, New Jersey, 2009.
- [25] L. Cao. "Support vector machines experts for time series forecasting", *Neurocomputing*, 51, pp.321-329, doi:10. 1016/S0925-2312(02) 00577-5.
- [26] F. Sönmez, Ş. Bülbül. "Intelligent Software Model Design for Estimating Deposit Banks Profitability with Soft Computing Techniques", *Neural Network World*, pp.319-345, 2015, doi: 10.14311/NNW.2015.25.017.
- [27] D. Altaş, A. M. Çilingirtürk, V. Gülpınar. "Analyzing the process of the artificial

- neural networks by the help of the social network analysis”, *New Knowledge Journal of Science*. 2(2), pp.80–91, 2013.
- [28] B. Yıldız, S. Akkoç. “Banka Finansal Başarısızlıklarının Sinirsel Bulanık Ağ Yöntemi ile Öngörüsü”, *BDDK Bankacılık ve Finansal Piyasalar*, 3(1), pp.9-36, 2009.
- [29] L. A. Zadeh. “The Roles of Fuzzy Logic and Soft Computing in the Conception, Design and Deployment of Intelligent Systems”, *BT Technology Journal*, 14(4), pp.32-36, 1994.
- [30] T. K. Kohonen. “The self-organizing map”, *Proceedings of the IEEE*, 78 (9), pp.1464–1480, 1990.
- [31] T. K. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, J. Honkela, V. Paatero, A. Saarela. “Self Organization of a Massive Document Collection”, *IEEE Transactions on Neural Networks*, 11(3), pp.574–585, 2000.
- [32] Bullinaria, J. A. Introduction to neural networks. *University of Birmingham*, UK, 2004.
- [33] L. Yang, Z. Ouyang, Y. Shi. “A Modified Clustering Method Based on Self-Organizing Maps and Its Applications”, *Procedia Computer Science*, 9, pp.1371-1379, 2012, doi: doi.org/10.1016/j.procs.2012.04.151.
- [34] J. A. Kangas, T. K. Kohonen, J. T. Jorma. “Variants of self-organizing maps”, *IEEE transactions on neural networks*, 1(1), pp.93-99, 1990.
- [35] N. R. Pal, J. C. Bezdek, E. C. K. Tsao. “Generalized clustering networks and Kohonen's self-organizing scheme”, *IEEE transactions on Neural Networks*, 4 (4), pp.549-557,1993.
- [36] B. Hammer, T. Villmann. “Generalized relevance learning vector quantization”, *Neural Networks*, 15(8–9), 1059-1068, 2002, doi: 10.1016/S0893-6080(02)00079-5.
- [37] T. M. Martinetz, S. G. Berkovich, K. J. Schulten. “Neural-gas network for vector quantization and its application to time-series prediction”, *Neural Networks, IEEE Transactions on*, 4(4), pp.558-569, 1993, doi: 10.1109/72.238311.
- [38] K. A. Smith. Introduction to neural networks and data mining for business applications. *Eruditions Publishing*, Melbourne, 1999.
- [39] D. Swagatam, D. Ajith, K. Amit, “Automatic kernel clustering with a multi-elitist particle swarm optimization algorithm”, *Pattern: Recognition Letters*, 29(5), pp.688–699, 2008.
- [40] J.Z. Lei, A.A. Ghorbani, “Improved competitive learning neural networks for network intrusion and fraud detection”, *Neurocomputing*, 75 (1), 135-145, 2012, doi: 10.1016/j.neucom.2011.02.021.
- [41] T. Fawcett, “ROC Graphs: Notes and Practical Considerations for Data Mining Researchers”, *Tech. Rep. HPL-2003-4, HP Labs*, 2003
- [42] A. Mitrokotsa, N. Komninos, C. Douligeris. “Intrusion Detection with Neural Networks and Watermarking Techniques for MANET,” *IEEE computer society*, pp.1-10, 2008.
- [43] W. Wanga, H. Wang, B. Wang, Yaping Wang, Jiajun Wang. “Energy-aware and self-adaptive anomaly detection scheme based on network tomography in mobile ad hoc networks,” *Information Sciences*, 220, pp.580–602, 2013.
- [44] G. M. Afify, A. E. Bastawissy, O. M. Hegazy. “A hybrid filtering approach for storage optimization in main-memory cloud database”, *Egyptian Informatics Journal*, 16(3), pp.329-337, 2015, doi: 10.1016/j.eij.2015.06.007.
- [45] A. T. Kabakus, R. Kara. “A performance evaluation of in-memory databases”, *Journal of King Saud University - Computer and Information Sciences*, 29(4), pp.520-525, 2017, doi:10.1016/j.jksuci.2016.06.007.
- [46] T. Lahiri, M. A. Neimat, S. Folkman. “Oracle TimesTen: An In-Memory Database for Enterprise Applications”, *IEEE Data Eng. Bull.*, 36(2), pp.6-13, 2013.
- [47] P. Jaroslav. “NoSQL databases: a step to database scalability in web environment”, *International Journal of Web Information Systems*, 9(1), pp.69-82, 2013.
- [48] P. Chao, D. He, S. Sadiq, K. Zheng, X. Zhou. "A performance study on large-scale data analytics using disk-based and in-memory

- database systems," *2017 IEEE International Conference on Big Data and Smart Computing (BigComp)*, Jeju, pp. 247-254, 2017, doi: 10.1109/BIGCOMP.2017.7881706
- [49] Y. Wang, G. Zhong, L. Kun, L. Wang, H. Kai, F. Guo. "The Performance Survey of in Memory Database", *2015 IEEE 21st International Conference on Parallel and Distributed Systems (ICPADS)*, Melbourne, Australia, pp.815-820, 2015, doi: 10.1109/ICPADS.2015.109.
- [50] S.-Y. Huang, R.-H. Tsaih, F. Yu. "Topological pattern discovery and feature extraction for fraudulent financial reporting", *Expert Syst. Appl.*, 41(9), pp.4360–4372, 2014.
- [51] P. C. González, J.D. Velásquez. "Characterization and detection of taxpayers with false invoices using data mining techniques", *Expert Systems with Applications*, 40 (5), pp.1427–1436, 2013.
- [52] S. Jha, M. Guillen, J.C. Westland. "Employing transaction aggregation strategy to detect credit card fraud", *Expert Systems with Applications*, 39, pp.12650–12657, 2012.
- [53] D. Olszewski. "A probabilistic approach to fraud detection in telecommunications", *Knowledge-Based Systems*, 26, pp.246–258, 2012.
- [54] V. D. Kumar, S. Radhakrishnan. "Intrusion detection in MANET using Self Organizing Map (SOM)", *2014 International Conference on Recent Trends in Information Technology*, Chennai, 2014, doi: 10.1109/ICRTIT.2014.6996118.
- [55] W. Khreich, E. Granger, A. Miri, R. Sabourin. "Iterative Boolean combination of classifiers in the ROC space: An application to anomaly detection with HMMs", *Pattern Recognition*, 43 (8), pp.2732-2752, 2010, doi: 10.1016/j.patcog.2010.03.006.
- [56] G. Kim, S. Lee, S. Kim. "A novel hybrid intrusion detection method integrating anomaly detection with misuse detection", *Expert Systems with Applications*, 41(4), pp.1690-1700, 2014.