IET Information Security

Research Article

IET Journals
The Institution of Engineering and Technology

C-NSA: a hybrid approach based on artificial immune algorithms for anomaly detection in web traffic

ISSN 1751-8709
Received on 26th October 2019
Revised 17th April 2020
Accepted on 8th June 2020
E-First on 30th June 2020
doi: 10.1049/iet-ifs.2019.0567
www.ietdl.org

Emre Dandıl[1] ✉

[1]*Department of Computer Engineering, Faculty of Engineering, Bilecik Seyh Edebali University, Bilecik, Turkey*
✉ *E-mail: emre.dandil@bilecik.edu.tr*

Abstract: Security vulnerabilities in web traffic can directly lead to data leak. Preventing these data leaks to a large extent has become an important problem to solve. Besides, the accurate detection and prevention of abnormal changes in web traffic is of great importance. In this study, a hybrid approach, called C-NSA, based on the negative selection algorithm (NSA) and clonal selection algorithm (CSA) of artificial immune systems for the detection of abnormal web traffic on the network is proposed and a user-friendly application software is developed. The real and synthetic data in the Yahoo Webscope S5 dataset are used for web traffic and the data are split into windows using the window sliding. In the experimental studies, the abnormal web traffic data is detected by monitoring the changes in the number of activated detectors in the C-NSA. It is observed that the average accuracy performance of finding anomalies in real web traffic data is 94.30% and the overall classification accuracy is 98.22% based on proposed approach. In addition, false positive rate of the proposed approach using C-NSA is obtained as 0.029. In addition, the results in synthetic web traffic data using C-NSA are achieved as average 98.57% classification accuracy.

1 Introduction

The Internet is a world-wide broadcasting capability, a mechanism used for information dissemination, and a medium for collaboration and interaction between individuals and computers regardless of geographic location. The Internet represents one of the most successful examples of the benefits of sustained investment and commitment to research and development of information infrastructure for millions of end users [1]. Internet technologies, which we have been actively using in our lives for many years, have faced a great development process and have gained a place in every area of our lives in particular. These technologies have entered our lives from private life to many fields in recent years [2]. With the widespread use of the Internet, various types of attacks are performed on networks. To solve such issues, researchers have carried out different types of studies on the network based on anomaly detection [3].

The internet has become embedded in every aspect of our day-to-day lives with the development of TCP/IP and many developments in the internet technologies. Therefore, the importance of information security and sustainability has gained significant value. It is now not enough to ensure security in the real environment, but also to ensure the security of the virtual environment [4]. The issue of to what extent the security precautions in the virtual environment has become measurable with the diversity and innovation of attacks on internet technologies. Today, cyber-attacks are employed in a more qualified way than the past and almost without a trace. In addition, wars from intelligence to military fields have been carried to cyber environment. Therefore, countries have started to make investments in internet technologies which provide more effective results with less money. Prevention of attacks has become a significant issue for the protection and sustainability of critical data.

Internet Security Threat Report [5] released by the Symantec in 2019 analysed data from global network. It is seen that events are recorded from 123 million attack sensors in more than 157 countries worldwide and 142 million threats are blocked daily. Although these threats are tried to be prevented, the ideas that in fact it is not possible to avoid these threats completely comes to prominence. Today, especially in the development of new attack types, determining the exact effects of these attacks and the necessary response is required a certain time. However, the time lost in this process leads to vulnerability of the system. At this point, detection of attacks and foreseeing abnormal situations on the data become important for network systems and web traffic. Considering that there are so many attackers today and what can be done by those who want to benefit from the vulnerability of the systems, it can be realised how important it is to detect the attack traffic, that is, the abnormal traffic. Abnormal traffic can damage the network as a result of various attacks.

Anomaly detection has become a vital component of any network in today's internet. Network traffic anomalies, which range from malicious unexpected attacks to network attacks such as denials-of-service and network scans, can have serious detrimental effects on the performance and integrity of the network [6]. The continuous arising of new anomalies and attacks create a continuous challenge to cope with events that put the network integrity at risk. Moreover, the complicated nature of traffic causes an increase in the protocol number, and complicates anomaly detection system's task. Although these methods have the ability to adapt to the changes in normal behaviour, most of the anomaly detection methods consider the detection methodology as static [7]. In addition, most network anomaly detection systems proposed so far has employed anomaly detection based on misuse signature-based detection methods [8]. Similar types of approaches are unable to detect and characterise unknown anomalies. As systems are monitored more closely and the reactions to the attacks are getting more and more attention, traditional rule-based systems for raising alerts become insufficient. Therefore, anomaly detection techniques based on machine learning have to be considered in order to make monitoring systems more dynamic and adaptive [9].

The common types of network attacks are denial of service, probe, user to root and remote to user [6]. Web traffic has different characteristics depending on the type of service provided, user connection patterns and irregularity in data distribution. Anomaly traffic detection in data classification can be conducted by different ways as from a point or as collective [4]. Today, the detection of these attacks and the measures to be taken are still being discussed. In this respect, it is really important to classify the data and determine whether the data is really harmful. Therefore, firstly detecting and then taking action procedures are important to avoid any data loss of the company and/or person. There are various solutions to prevent this loss. These solutions make it much easier to secure data and implement corporate policies.

There are some studies in the literature for the detection of abnormal traffic on the network [10–12]. Several methodologies and data classification techniques are used to detect abnormal traffic in network data [13, 14]. This problem is generally evaluated by classifying the signal windows by feature extraction from network traffic data. However, there are few studies to detect the abnormal situation in traffic data using the negative selection algorithm (NSA) of artificial immune algorithms. Meanwhile, in our previous preliminary study [6] performed on Yahoo Webscope S5 dataset [15], a method based on the NSA of artificial immune system (AIS) for the detection of abnormal web traffic on the network is proposed in real web traffic data only. The contributions of the work are summarised as follows:

(i) In this study, we propose a hybrid approach (C-NSA) based on NSA optimised using clonal selection algorithm (CSA) of AISs for anomaly detection in web traffic.
(ii) Consequently, the threshold values used for the manually selected by NSA are determined automatically by optimising with CSA.
(iii) In the study, Yahoo Webscope S5 [15] dataset is used for the experimental studies. As a result, the performance of detecting abnormal traffic in the network is increased with the determination of the most suitable threshold values using C-NSA.
(iv) In addition, user-friendly software is developed for experimental results. The traffic values of the network data in the time series are used by the developed software based on C-NSA.
(v) In the study, detection of abnormal web traffic occurring in the network and detection of at which time steps abnormal traffic values are detected with high accuracy successfully.
(vi) When the studies in the literature are investigated, it is seen that the detection procedures for anomaly detection in network traffic are handled separately and various methods are proposed for each procedure. On the other hand, the proposed method enables performing all procedures with high performance by using C-NSA.

2 Related work

There has been an increase in both targeted and automated attacks in recent years and sophisticated and layered defence mechanisms are needed in order to mitigate these threats. A two-line defence mechanism is a common example. First line of defence is typically deployed on the network perimeter. The structure that is composed of various internet-facing devices including routers and firewalls creates this line. These are configured to block easily-identifiable illegitimate traffic. The second line of defence is an intrusion detection system (IDS). IDS is a system built to analyse events within the network for the detection of ongoing malicious activities. Evidence of attacks is reported to the network administrators according to the logs, so they can react accordingly [16]. IDSs provide an efficient defence to avoid network attacks on the Internet. IDS also could be detected different types of attacks on network where the traditional firewall systems cannot overcome well [17].

The anomaly-based intrusion detection technique assumes that malicious activities are significantly different from expected behaviour, and react to these differences. The incoming events are analysed to check whether they deviate from the normal ones. Anomaly-based systems support detection of unknown and novel attacks and can also be trained to overcome the problems caused by the security gaps. Anomaly-based components can be modelled by using different machine learning techniques [18].

Detection of abnormal traffic data on web servers is often considered as a time series problem. This problem is assessed by extracting features on signal windows and classifying them [19]. However, since web traffic data do not have a general characteristic pattern structure, the solution methods also differ. There are many studies in the literature for the classification of abnormal data on the network. In their study, Münz *et al.* [20] conducted anomaly test in the network using K-means clustering algorithm. They calculated the cluster centroids by analysing the statistical properties of the real data. Moreover, they presented a novel flow-based anomaly detection scheme based on the K-means clustering algorithm. Thill *et al.* [21] compared several online anomaly detection algorithms on Yahoo Webscope S5 dataset for anomaly detection. They proposed an innovative, online distance based anomaly detection algorithm. The results obtained from regression analysis showed that the anomaly detector was quite successful compared to other anomaly detectors. Kim and Cho [4] proposed a C-LSTM neural network for effectively modelling the spatial and temporal information contained in traffic data, which is a one-dimensional time series signal. Experiments demonstrated that the proposed C-LSTM method can extract more complex features by combining a convolutional neural network (CNN), long short-term memory (LSTM) and deep neural network. The C-LSTM method achieved successful anomaly detection performance for web traffic data, even for very similar signals that were previously considered to be very difficult to classify. In the study conducted by Akbal and Ergen [22], attack detection in wireless networks was approached from a different perspective using AIS. In the study, control operation was performed over access point instead of performing detection on all users. As a result of long-term observations on the network, it was concluded that AIS successfully acted and when the working time was longer, the system was more successful. Dutt *et al.* [23] demonstrated that malicious attacks on the network system could be detected efficiently by using the AIS. Certain servers were tried to be infected by malicious software such as virus or worms. Kim and Bentley [24] investigated the role of negative selection in AIS for network intrusion detection. The study focused on the use of negative selection as a network traffic anomaly detector. Zhang *et al.* [25] examined the vulnerabilities of wireless networks and argued that intrusion detection must be included in the security architecture for mobile computing environment. Das *et al.* [26] conducted anomaly detection, which is a concept widely applied to a number of domains, based on anomaly-based IDS which can detect previously unknown attacks, which is important since new security vulnerabilities and attacks are constantly appearing. Moreover, deep learning based methods have recently been used for malware detection on network. In one of such studies, Bakhshinejad and Hamzeh [27] proposed a method for malware detection using parallel architecture of CNN.

In many research areas, hybrid approaches can be proposed by combining different techniques for the problem that good results cannot be obtained with only one method [28, 29]. In networks, hybrid approaches come to the fore especially with the help of optimisation techniques for processes such as anomaly and intrusion detection. In one of the studies proposed in this way, Ganapathy *et al.* [30] proposed fuzzy c-means clustering method based on immune genetic algorithm for intrusion detection. Selvi *et al.* [31] performed an intelligent agent and fuzzy swarm optimisation based routing algorithm for dynamic clustering on wireless sensor network. Similarly, in this study, a novel hybrid approach based on the NSA and CSA of AISs, called C-NSA, is proposed for the detection of abnormal web traffic on the network.

3 Proposed method

In this study, abnormal web traffic data was detected with the proposed method using C-NSA on Yahoo Webscope S5 [15] dataset as one-dimensional time series. The open block diagram showing the detection structure and flowcharts of the proposed method is presented in Fig. 1. As seen in Fig. 1, data is first obtained from the dataset. In the second step, this data is divided into a certain number of windows, and the detection procedure is applied to each window by sliding window procedure. After the training and test windows are determined with the NSA algorithm optimised by CSA, the abnormal web traffic data in each window are detected by the detectors activated at the last stage using C-NSA.

3.1 Web traffic dataset

Anomaly detection in web traffic data is often treated as a time series signal. In this study, the real and synthetic time series data of web traffic in the Yahoo Webscope S5 [15] dataset were used for anomaly detection on network. This dataset consists of four different classes such as A1, A2, A3 and A4 and totally there are

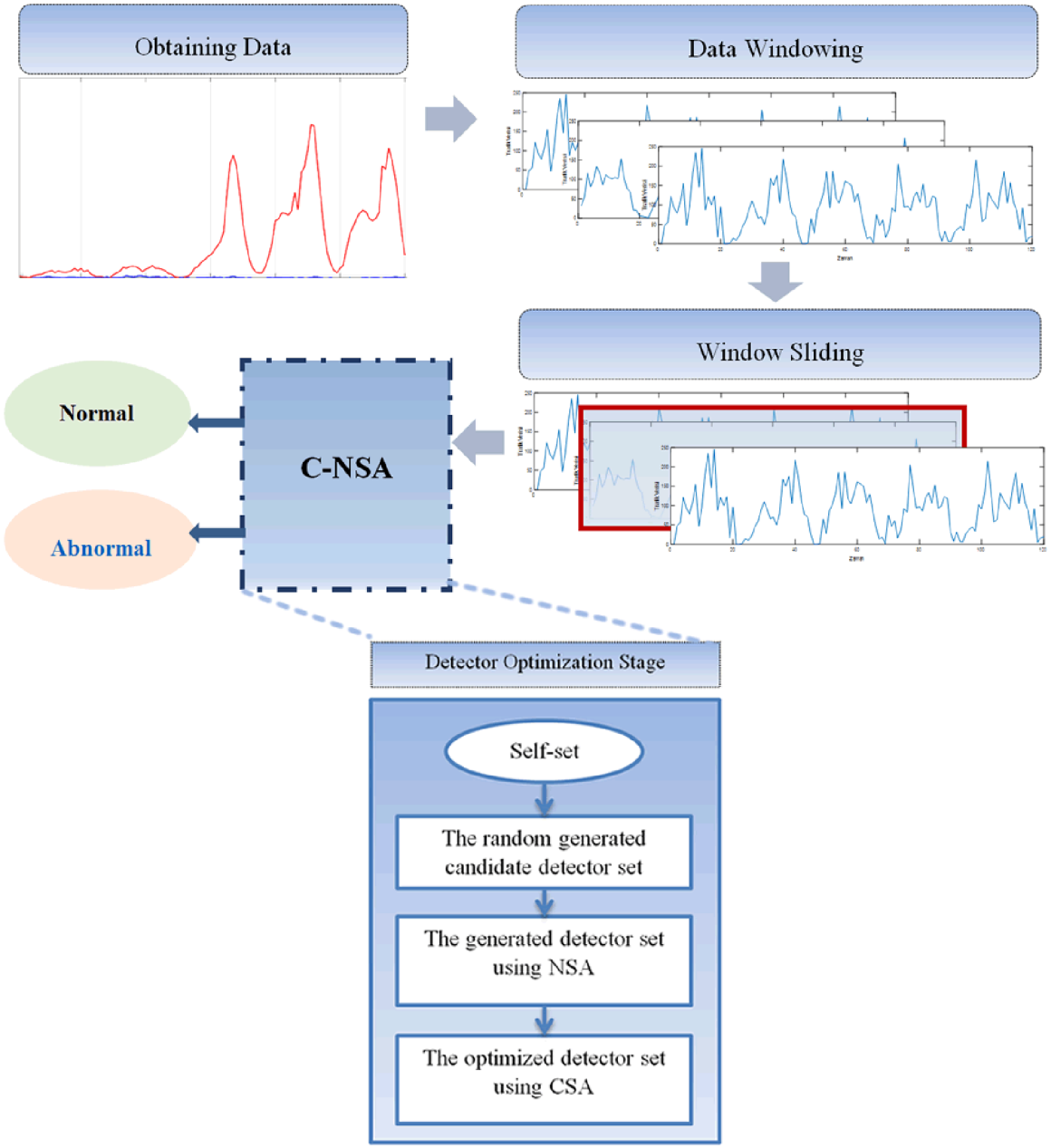

Fig. 1 *Detection structure of the proposed C-NSA approach. The architecture of the system comprises of four sub-sections such as obtaining data, data windowing, window sliding and C-NSA*

Table 1 Details of Yahoo Webscope S5 dataset

Class name	Real/synthetic traffic (R/S)	Total length	Anomaly in class
A1	R	94,866	1669
A2	S	142100	466
A3	S	168,000	943
A4	S	168,000	837

367 time-series signal patterns. The detailed information about the Yahoo Webscope S5 dataset was provided in Table 1. Each signal pattern contains an average of 1500 data points, and there are a total of 5,050,000 data points in four different classes. A1 class contains real data, while web traffic data in other classes contain synthetic data. In this study, A1 class which contains real web traffic data and A2 class which contains synthetic data web traffic was used in experimental studies because the signal attributes of both classes are equal and similar. The A1 class in the Yahoo Webscope S5 dataset consists of 67 real web traffic time series, while there are 100 synthetic web traffic time series in A2 class.

All data files (in A1 and A2 classes time series) are given in the form of CSV files with three columns and the values are as time stamp, traffic value and whether or not the anomaly (1 or 0). While the time stamp includes the time values at which the traffic occurs, the value is the part of the instantaneous data in the network and anomaly is the part that tells us whether or not the anomaly exists. A couple of signal pattern including abnormal web traffic data are denoted in Figs. 2*a* and *b* for A1 class and A2 class, respectively. The sharp rise values seen in the time series signals of both signals indicate abnormal web traffic.

3.2 Data windowing

A1 and A2 classes in the Yahoo S5 dataset are divided into 12 windows to detect abnormal data to detect abnormal web traffic data in signal. Each window has a time scale of 120 steps (points). The purpose of the windowing is that normal traffic values in window are needed to detect abnormal traffic. The windowed normal traffic values are used as training data, and abnormal traffic is detected by the experimental studies using the proposed approach, which is named as N-CSA. The partition of a sample signal pattern into windows, and areas with normal and abnormal

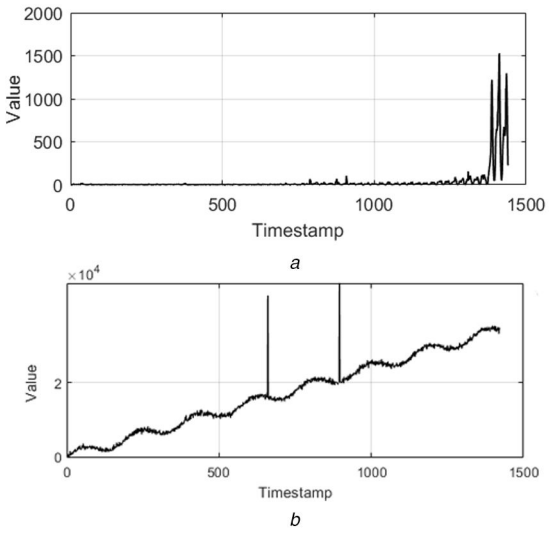

Fig. 2 *Couple of signal pattern samples in Yahoo Webscope S5 dataset with normal and abnormal web traffic data*
(a) Signal pattern for real web traffic time series in A1 class with abnormal data, ***(b)*** Signal pattern for synthetic web traffic time series in A2 class with abnormal data

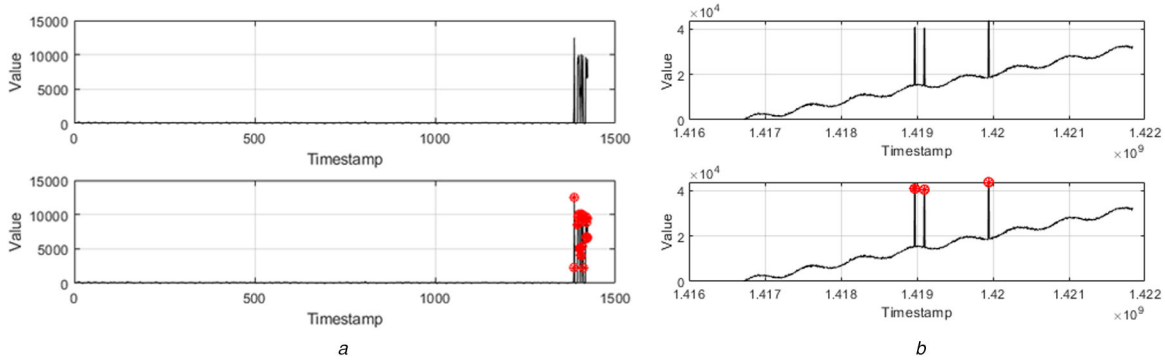

Fig. 3 *Normal and abnormal web traffic data in real and synthetic data*
(a) Real abnormal and normal web traffic data in time series of A1 class, ***(b)*** Synthetic abnormal and normal web traffic data in time series of A2 class

traffic data are shown in Fig. 3. In Figs. 3*a* and *b*, the dots in the red marker show abnormal web traffic data in real and synthetic time series signals, respectively.

3.3 Window sliding

The sliding window technique examines the most recent data points when a continuous time series stream is given. In addition, sliding window moves steps along the time axis as new measurements arrive. The main advantage of this technique is that it does not need to store the never-ending stream of data, but this also implies that measurements can only be considered for further data analysis as long as they are located within the current window. By segmenting the signal patterns obtained by time series into windows, new values are obtained by preserving the dimensions and important features of the time series data. The objective of dividing is to decrease the error in the original time series and to obtain the best data. In this approach, the entire signal is processed by sliding the window between the segmented signal windows. For various time series applications such as weather, finance and medical, window sliding method procedure is often used. Starting from the first value, windows partitions are specified. After selecting the first partition, the next partition is processed according to the given criteria. The process is repeated until all-time series data are divided. This method is intuitive and simple. The aim of this method is to reduce the overall approximation of given specific amount of information [32]. In this study, a time series signal with web traffic data is divided into multiple windows and each window includes in 120 data points. First of all, foremost window is processed, and then estimated values are generated after by sliding to the next window. In each window, the algorithm is run and network training is performed. As a result, anomaly detection and attack areas in web traffic are detected using proposed approach. All the data points in a time series of A1 class and the whole window partitions with 120 data are denoted in Fig. 4.

3.4 Artificial immune systems

The human immune system protects the human body against harmful and previously unseen foreign cells using lymphocyte cells. The foreign cells are called antigens, such as bacteria and viruses. The AIS is designed for the computational system and inspired by the human immune system [33]. This system is applied to solving various problems in the field of information security, particularly in IDSs. Moreover, it contains many features of the human immune system, including diversity, error tolerance, dynamic learning, adoption and self-monitoring. It differentiates between 'self' (cells that are owned by the system) and 'non-self' (foreign entities to the system) as intrusions. Similarly, detectors similar to lymphocytes are deployed in computer system in order to prevent and report any malicious activities [34]. AIS can be broadly divided into two categories based on the mechanism. These categories are network-based models and population-based models. Besides, there are many hybrid models. These AIS's are built on algorithms which accept that there are interactions between antibodies and antigens as well as between antibodies and antibodies. Population-based models use negative or clonal selection as the method of generating and maintaining a population of detectors. These models are used to build some IDSs [35].

The immune system is capable of recognising, defining and responding to a wide range of different models. In addition, the immune system may vary between non-self-functioning and self-damaging cells [36, 37]. The immune system should be able to distinguish between self and non-self-antigen. Receptor molecules have roles in this procedure. These receptors are divided into two groups such as B and T cells [38]. These two types of cells are

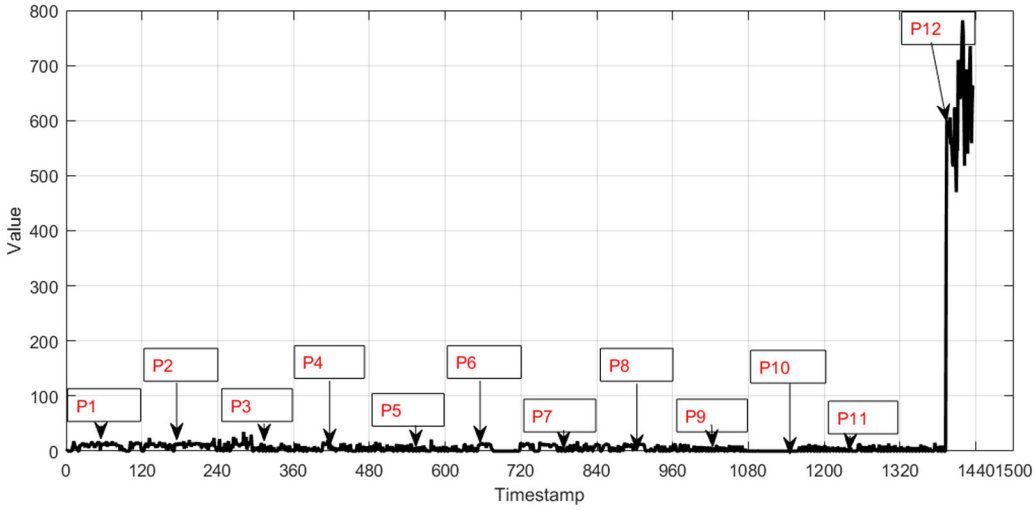

Fig. 4 *Entire signal pattern for windows with abnormal web traffic*

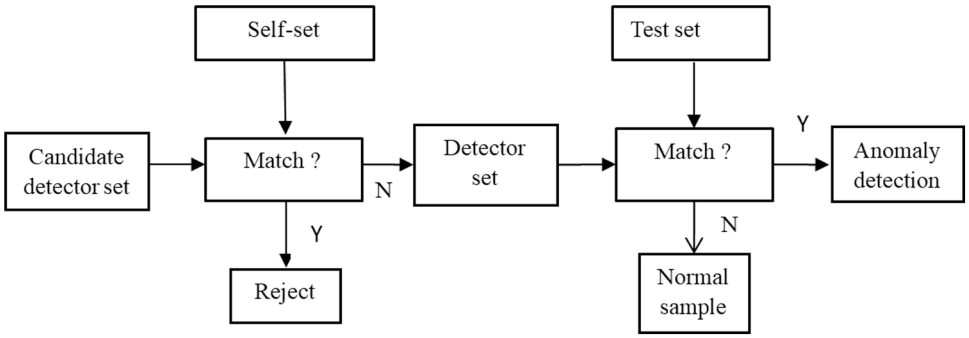

Fig. 5 *Flowchart of NSA*

actually quite similar, but they differ in how they recognise antigens and how they determine their roles. The human immune system plays an important role in the maturation of thymus (T) cells located behind the breastbone. During this maturation, all T cells are subtracted from the T cell population in identifying recognised antigens. This event is called negative selection. If a B cell encounters an antigen that is not identical to it, it multiplies and differentiates into memory and effector cells. This event is called clonal selection [38].

Today, when the studies in the literature are examined, many approaches has been developed and used in various studies such as pattern recognition, computer security, disease diagnosis, optimisation problems and dynamic programming [39]. AIS was developed based on the human immune system. When any harmful population penetrates our body, the immune system reacts to the model to detect abnormal conditions. Detection algorithms are developed by transferring these models developed on the immune system to the computer system. AIS has four main algorithms as NSA or positive selection algorithm, CSA, immune network models and antibody network model [40].

3.4.1 Negative selection algorithm: The recognition of the harmful organism entering the body in the immune systems of living organisms is performed by B and T cells. Of these cells produced in the bone marrow, T cells are subjected to a process called negative selection in the thymus. NSA was developed as a result of observing these cells [41]. When Forrest *et al.* [42] proposed the initial NSA, they used binary encoding to represent the normal and abnormal space. Later, a real-valuable approach was presented and a genetic technique was proposed to close the abnormal space and generate good detectors. Among the recent studies on NSAs, Ji and Dasgupta [43] proposed V-detector consisting of real values. Balachandran *et al.* [44] proposed a multi-shape detector NSA. The NSA is one of the most successful methods in AIS and typical applications include change detection, fault detection and network intrusion detection. Although there have been many successful applications of the NSA, some problems still exist in order to prevent the AIS and the NSA from being extensively applied [45]. The process steps of the NSA algorithm, which was developed with inspiration from this process, can be listed as follows:

(i) First, a self-set is determined from the dataset.
(ii) In the next step, the candidate detectors are generated randomly in defined number.
(iii) In third step, among the candidate detectors, the ones which match with the determined threshold in training set are removed. Non-matched candidate detectors with self-set are added to the detector set. The Euclidian distance measurement given in (1) was used to calculate the match between a candidate detector and the training (self-set) or test set. Where ϵ denotes threshold value between detector set and self-set, A represents the distance, l represents the data number, Ab represents the test or training set, and Ag represents the detector set. According to (2), if $E > 0$, the sample in the candidate detector set is added to the mature detector set. If $E <= 0$, these samples are rejected

$$A = \sqrt{\sum_{i=1} (Ab_i - Ag_i)^2} \tag{1}$$

$$E = \epsilon - A \tag{2}$$

(iv) In the test step, the test set is also created from the dataset. It is checked whether each element in the detector set is matched to the threshold value determined between each element in the test set. If there is a match ($E < 0$), an abnormal condition is detected, and if there is no match, it is normal.
(v) The process is ended by presenting the results of the test step.

The flowchart of the NSA, which generally shows the process flows, is shown in Fig. 5. NSA detector generation starts with a population of candidate detectors that mature through an iterative process.

```
input  : S = set of patterns to be recognized, nthe number of worst elements to
select for removal
output : M = set of memory detectors capable of classifying unseen patterns
begin
  Create an initial random set of antibodies, A
  For all patterns in S do
    Determine the affinity with each antibody in A
    Generate clones of a subset of the antibodies in A with the highest affinity.
      The number of clones for an antibody is proportional to its affinity
    Mutate attributes of these clones to the set A , and place a copy of the
      highest affinity antibodies in A into the memory set, M
    Replace the n lowest affinity antibodies in A with new randomly generated
    antibodies
  end
end
```

Fig. 6 *Pseudo-codes of CSA mechanism*

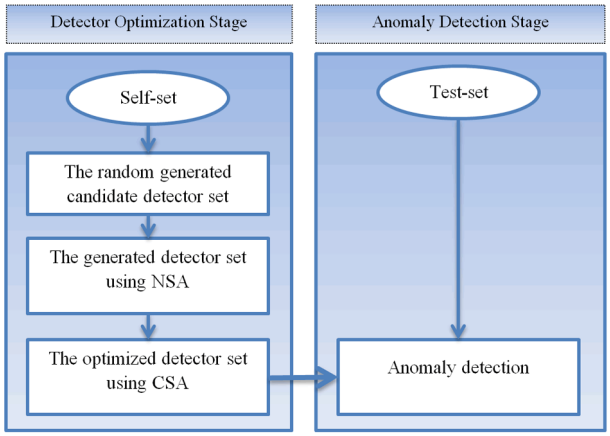

Fig. 7 *Clonal optimisation of NSA for anomaly detection in web traffic (C-NSA)*

3.4.2 Clonal selection algorithm: The clonal selection principle is used to explain the basic features of an adaptive immune response to an antigenic stimulus [46]. According to this principle, only cells that recognise antigens proliferate. The selected cells are subject to an affinity maturation process, which improves their affinity to the selective antigens [47, 48]. Fig. 6 shows the pseudo-codes and the operation of the CSA algorithm.

The processing steps of the CSA can be described as follows. A random set of antibodies is initially generated to show the set of antibodies A. Afterwards, a series of procedures is repeated. In the first of these processes, the affinities of each pattern in set A are determined according to the fitness criteria. As a second process, the patterns with the highest affinity from A set are cloned to form a subset. Here, the number of clones for each antibody in A set is proportional to its affinity criterion. In the third process, each pattern is mutated in A set. The patterns with the highest affinity are then determined and added to the M memory set. Finally, n antibodies with the lowest affinity in A set are replaced by new randomly generated antibodies. These processes are repeated until the stop criterion is provided, that is, until all patterns in the set S are recognised.

3.5 Optimisation of NSA using CSA

In this study, CSA was used during the production of the most suitable set of detectors matured with NSA. This approach was named as C-NSA. Here, the C-NSA performs the function of optimally covering the training signal of the generated detector set by adaptively determining the training-detector and training-test threshold values.

It is seen that the detector set generated by classical NSA algorithm does not fully coincide with the self-set. That is, a weaker set of detectors is formed. This restrictive situation may lead to low performance in anomaly detection stage in web traffic data. Therefore, in this study, the detector set generated by NSA was optimised with CSA to obtain the most suitable detector set. In the optimisation procedure with CSA, the function given in (3) was used to calculate fitness (f). Accordingly, the suitability value f for CSA is calculated according to whether the value of a detector is equal to or greater than the difference between the average of the training set S and the T test set

$$f = \frac{1}{n}\sum_{i=1}^{n} S(i) - \sum_{i=1}^{n} T(i) \tag{3}$$

Here, f represents the fitness value for CSA, S represents the training set, T represents the test set and n indicates the number of patterns in the training and test sets. Clonal optimisation of NSA for anomaly detection in web traffic (C-NSA) is denoted in Fig. 7.

4 Developed application and experimental results

4.1 Developed application

In this study, the user interface with Matlab GUI designed for detecting abnormal traffic in web traffic. In the developed software, the data which are determined as the normal value that is windowed in the training data section are loaded first. In the test section, the data which are determined to be abnormal and windowed are loaded. After the AIS parameters (training-detector threshold, test-detector threshold etc.) are determined in the conducted experiments using C-NSA, the anomaly in the web traffic is determined by the software according to the selected training and test data. The percentage of the errors occurring in the abnormal traffic data on the network is determined according to the number of activated detectors at this stage. Experimental studies were conducted on the time series data contained in the A1 and A2 classes of Yahoo Webscope S5 dataset to show accuracy of the application. In these experiments, the first 120 data was named as $P1$, the second 120 data was named as $P2$ and others continue in the same way. Thus, a total of 12 windows, P1−P12, are created for each time series signal. There is a total of 120 data in each window and abnormal web traffic data in each window is detected using the sliding window. For these experiments, optimal results were calculated with training threshold and test threshold values using C-NSA.

Experimental studies on Yahoo Webscope dataset were performed with signal patterns belonging to real data in A1 class

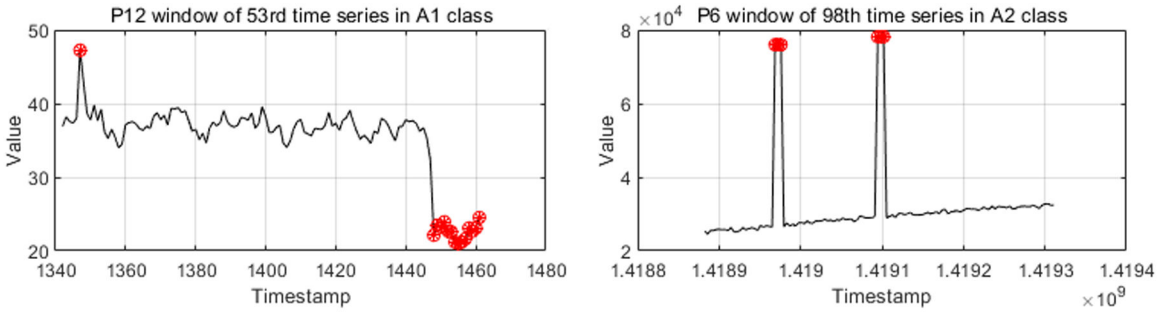

Fig. 8 *The windows of abnormal time series for real and synthetic data*
(a) P12 window with abnormal real web traffic data in 53rd web traffic signal of A1 class, ***(b)*** P6 window with abnormal synthetic web traffic data in 98th web traffic signal of A2 class

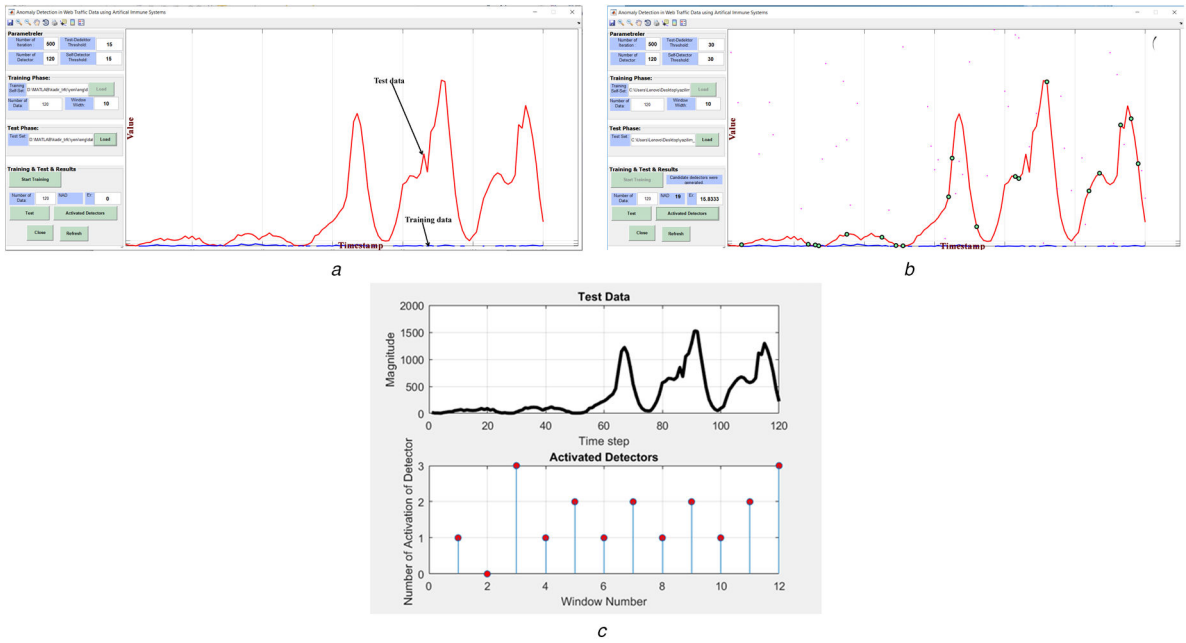

Fig. 9 *Detection of abnormal web traffic on the developed software using the proposed method (C-NSA)*
(a) Loading the training and the test data to application software for experimental studies, ***(b)*** Detection of abnormal web data on the signal window by C-NSA, ***(c)*** Activated detectors in the signal window with abnormal web data

and synthetic data in A2 class. The abnormal traffic data window of 53rd signal pattern of 12th (P12) window in A1 class is shown in Fig. 8*a* and the abnormal traffic data window of 98th signal pattern of 6th (P6) window in A2 class is denoted in Fig. 8*b*. As seen in these figures, the selected windows for web traffic signal in A1 and A2 classes, that is, the data with the abnormal traffic and the data used as test data, which is marked with a red circle, indicates the abnormal data.

The first 120 data (P1) of the 42th signal were specified as the training data for the experimental studies and the last 120 data (P12) were specified as the test data, and they were loaded into the developed application, which is presented in Fig. 9*a*. The training data of P1 window and test data of P12 window are specific to this experimental study only. The selected windows may change in other experiments. In this test process, the number of training detector thresholds and the number of test detector thresholds were selected as 15 for both them. However, the number of training detector thresholds and the number of test detector thresholds were determined as 30 by CSA for both of them, as seen in the form interface of Figs. 9*a* and *b*. For experimental studies, abnormal data detected by the C-NSA in 42th signal is presented in Fig. 9*b*. It is seen that the errors occurring in the abnormal traffic data in the network change in proportion to the number of activated detectors at this stage. Fig. 9*c* shows the number of activated detectors and their position in the signal window for test signal with abnormal traffic in the experimental study. Activated detectors during the analysis of the data in the network are the main indicators of the anomaly. Activated detectors also indicate at which time steps the anomaly data is anomaly. As can be seen, the number of activated detectors in areas where the anomaly is high on the signal increases.

4.2 Experimental results

All applications were performed on the application with user interface created with MATLAB software. All experimental studies were carried out using a computer which is described in Table 2. In the study, the training set is first determined from a signal window. After the training phase is completed, anomaly detection was performed with C-NSA in test signals with anomaly. The number of iterations was determined as 500 and the number of detectors was determined as 50 in all experiments. However, the training-detector threshold and test-detector threshold values were determined adaptively by CSA using the weights of the used training and test signals.

In the study, experimental studies were conducted on 7th, 22nd, 25th, 40th, 42nd, 58th and 67th web traffic signal patterns of A1 class and in the web traffic data of 1st, 10th, 15th, 26th, 47th, 89th and 98th time series in A2 class in the Yahoo Webscope S5 data. Each of these signal was divided into 12 different windows of 120 data points during the test since as seen in Table 3, the most successful result in detecting anomaly in 58th web traffic data in different window widths was obtained when a window is 120

Table 2 Configuration information of the computer used for experimental studies

Hardware	Configurations
processor (CPU)	Intel Core i7-7700 K @ 2.8 Ghz (8 CPUs)
memory (RAM)	16 GB (DDR4 2400 Mhz)
mainboard	ASUS X580VD
GPU	NVIDIA GeForce GTX 1050 (4 GB)
harddisk driver (HDD)	256 GB SSD

Table 3 Effect of window width on anomaly detection in web traffic data for the 58nd time series in A1 class (FID = file number, trd = training data, ted = test data, nop = number of points, noa = number of anomaly, noad = number of activated detectors, noda = number of detected anomaly, acc = accuracy, trdt = training-detector threshold, tedt = test-detector threshold)

	trd	ted	nop	noa	noad	noda	acc	trdt	tedt
FID: 58	P1	P72	20	20	18	15	75.0	20	40
	P1	P36	40	40	20	34	85.0	20	40
	P1	P20	72	43	25	38	88.37	20	40
	P1	P12	120	43	22	41	95.34	20	40
	P1	P16	90	43	27	39	90.7	20	40
	P1	P10	144	43	23	40	93.02	20	40

Table 4 Results obtained on the windows with abnormal web data for the 42nd data and the used parameters based on C-NSA

	trd	ted	nop	noa	noad	noda	acc	trdt	tedt
FID: 42	P1	P2	120	0	0	0	100	30	30
	P1	P3	120	0	0	0	100	30	30
	P1	P4	120	0	0	0	100	30	30
	P1	P5	120	0	0	0	100	30	30
	P1	P6	120	0	0	0	100	30	30
	P1	P7	120	0	0	0	100	30	30
	P1	P8	120	0	0	0	100	30	30
	P1	P9	120	0	0	0	100	30	30
	P1	P10	120	0	0	0	100	30	30
	P1	P11	120	0	0	0	100	30	30
	P1	P12	120	44	18	41	93.18	30	30

Table 5 Confusion matrix of the actual and predicted normal and abnormal web traffic data for 42nd data using C-NSA

Confusion matrix	Predicted normal	Predicted abnormal	Total sample
actual normal	76 (TP)	0 (FN)	76
actual abnormal	3 (FP)	41 (TN)	44
total sample	79	41	120

points. Therefore, in all tests, the window width was determined to consist of 120 points.

One of the windows was determined as self-set and the window with anomaly was determined as non-self-set. The results obtained from tests performed for the detection of normal and abnormal traffic for 42nd pattern were denoted in Table 4. As seen in the table, training-detector threshold value was determined as 30 and the threshold of test-detector was determined as 30 by CSA. In addition, P1 was selected for the training set and P12 was selected for the test set. In the test procedure conducted with abnormal traffic, the total number of anomalies was 44, the number of the activated detectors was 18, the number of correct anomalies was 41 and the accuracy rate was found as 93.18%. Besides, as can be seen in Table 4, when a signal window with no anomaly (training set) and another signal window with no anomaly were processed, there was no abnormality in the test result and all data points were detected as normal with 100% performance.

In the evaluation of anomaly detection performance with application, the detection performance of detectors with anomaly in traffic is shown in confusion matrix in Table 5. As can be seen, all of the 76 normal data in the P12 window of 120 web traffic data were found to be normal (true positive, TP), while 41 of the 44 abnormal data were found to be abnormal (true negative, TN). In addition, since all data in the normal category were correctly classified, there were no abnormal (false negative, FN) data, but three of the abnormal data were classified as normal (false positive, FP). The performance evaluation of the classification process is calculated using the accuracy metric shown in (4). Therefore, the classification accuracy was found as 97.50% according to the accuracy table in Table 5

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{4}$$

The results obtained for the anomaly detection accuracy rates and other parameters of the remaining data 17th, 22nd, 25th, 40th, 67th and 58th from the real data in class A1 in the dataset were presented in Table 6.

As can be seen, the accuracy of detection in all data is greater than 93%. Furthermore, it can be easily concluded that the training

Table 6 Anomaly detection accuracy rates and parameter analysis in test data using C-NSA (FID = file number, trd = training data, ted = test data, nop = number of points, noa = number of anomaly, noad = number of activated detectors, noda = number of detected anomaly, acc = accuracy, trdt = training-detector threshold, tedt = test-detector threshold)

FID	trd	ted	nop	noa	noad	noda	acc	trdt	tedt
17	P1	P12	120	79	20	74	93.67	10	65
22	P1	P12	120	63	21	59	93.65	35	35
25	P1	P12	120	42	24	40	95.20	15	80
40	P1	P10	120	76	23	71	93.42	15	15
67	P11	P12	120	23	16	22	95.65	20	60
58	P1	P12	120	43	22	41	95.34	20	40

Table 7 Confusion matrix of classification performance for test windows using C-NSA

FID	Test window no.	Total data	TP	FN	FP	TN	Acc, %	FPR
17	P12	120	43	0	3	74	97.50	0.038
22	P12	120	58	0	3	59	97.50	0.048
25	P12	120	79	0	1	40	99.17	0.024
40	P10	120	46	0	3	71	97.50	0.041
67	P12	120	97	1	0	22	99.17	0.000
58	P12	120	78	0	1	41	99.17	0.023

Table 8 Summary table of the results obtained with experimental studies in real time series of A1 class (TN = test no, FID = file number, AFR:NSA = anomaly finding rate using NSA,OCA:NSA = overall classification accuracy using NSA,AFR:C-NSA = anomaly finding rate using C-NSA,OCA:C-NSA = overall classification accuracy using C-NSA)

TN	FID	AFR: NSA, %	OCA: NSA, %	AFR: C-NSA, %	OCA: C-NSA, %
1	42	88.64	92.50	93.18	97.50
2	58	90.70	92.50	95.34	99.17
3	17	91.14	90.83	93.67	97.50
4	22	88.89	91.67	93.65	97.50
5	25	92.86	92.50	95.20	99.17
6	40	90.78	92.50	93.42	97.50
7	67	91.30	92.50	95.65	99.17
average		90.62	92.14	94.30	98.22

and test windows vary in these tests and that the training-detector and training-test threshold values determined by the CSA are adaptively determined according to the structure of the data.

Table 7 also shows the classification performance accuracy of normal and abnormal data in the selected test window for 17th, 22nd, 25th, 40th, 42nd, 58th and 67th real web traffic time series signals in A1 class. As can be seen from this table, it can be concluded that the proposed C-NSA based detection method reached a classification accuracy of 98% for detection of anomalies in web traffic data. In addition, false positive rate (FPR), also known fall out, is one of the important criteria used to measure the performance of the system in pattern recognition problems. FPR is calculated as the proportion of negative patterns incorrectly identified as positive patterns in the training or test data and denoted in (5). In Table 7, when the FPR measurement results are examined, it is seen that the average FPR value is reached to 0.029, as well. It can be deduced from this that the number of false alarms is low when finding abnormal traffic in the system's web traffic data

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{5}$$

Anomaly finding rate (AFR) defines the correct detection rate of anomalies in the time series web traffic data used in the testing process. Overall classification accuracy (OCA) represents the accuracy rate in abnormal or normal classification of abnormal and normal web traffic data in the time series web traffic data used in the testing process. Besides, while AFR also shows the Acc metric presented in Table 4, OCA gives the Acc metric results calculated in Table 7 based on the confusion matrix in Table 5. The results obtained for both NSA and C-NSA on each data are presented in Table 8. As can be seen from this summary table, with the proposed C-NSA method, the performance for detection of anomalies in a web traffic data (AFR) was found as 94.30% on average and the performance of the OCA rate was found as 98.22% on average. However, the average OCA and AFR values calculated using conventional NSA were 90.82 and 92.09%, respectively. Thus, it can be said that the proposed method is successful in detecting anomalies in web traffic data with high performance when comparing results obtained with using conventional NSA.

In this study, in experimental studies, testing procedures were carried out using C-NSA approach for anomaly detection in synthetically produced web traffic data as well as real web traffic data since the attributes of the time series containing web traffic data in both classes are of the same structure. While there is a total of 1669 abnormal data in A1 class consisting of real web traffic data, there are 466 abnormal data in A2 class consisting of synthetic web traffic data. Although there are more abnormal data in the time series consisting of A1 class web traffic data and generally located in the last window of the time series, there is less abnormal web traffic data in the class A2 which is produced synthetically and is seen in different windows of the time series. In the study, as shown in Table 9, anomaly detection was performed with the proposed C-NSA approach in the web traffic data of 1st, 10th, 15th, 26th, 47th, 89th and 98th time series in A2 class. In addition, it is seen in Table 9 the number of abnormal data in web traffic in each time series on which experimental work is carried out. According to the experimental studies conducted in synthetic web traffic data in A2 class, the average value for AFR was 96.82% and the average value for OCA was 98.57% using the C-NSA method. The average OCA and AFR values obtained as a result of experimental studies on synthetic web traffic data in A2 class show that the C-NSA approach proposed in this study is successful in detecting anomalies in web traffic data when comparing with the results obtained using conventional NSA.

Table 9 Summary table of the results obtained with experimental studies in synthetic time series of A2 class

TN	FID	noa	AFR: NSA, %	OCA: NSA, %	AFR: C-NSA, %	OCA: C-NSA, %
1	1	4	100	94.17	100	99.17
2	10	4	100	91.67	100	100
3	15	1	100	91.67	100	100
4	26	9	66.67	90.83	88.89	97.50
5	47	9	88.89	93.33	100	98.33
6	89	9	100	94.17	100	96.67
7	98	9	77.78	90.83	88.89	98.33
average:			90.48	92.38	96.82	98.57

Table 10 Comparison of the proposed approach (C-NSA) with similar studies in the literature

The studies	Dataset	Method	Acc.	FPR
Kim and Cho [4]	Yahoo Webscope S5 [15]	C-LSTM	98.60	*NA*
Garg *et al.* [49]	Yahoo Webscope S5 [15]	BFA-PDBSCAN	98.90	0.026
Dandıl and Ilhan [6] (our preliminary work)	Yahoo Webscope S5 [15]	NSA	95.38	*NA*
proposed approach (C-NSA)	Yahoo Webscope S5 [15]	C-NSA	98.22	0.029

Although it is difficult to compare this study with the literature in terms of many parameters, it can be evaluated in terms of accuracy and false alarm with studies using the same dataset. Table 10 presents a comparison of similar studies in the literature using the same dataset with the hybrid approach C-NSA proposed in this study on Yahoo Webscope S5 dataset in terms of FPR and Acc. As can be seen from Table 10, the results in terms of FPR and Acc achieved by the C-NSA approach proposed in this study are similar to results obtained by some studies in the literature on the same dataset. In this study, more successful results with 98.22% Acc using C-NSA were obtained when comparing with the results in our previous study [6] using NSA. Besides, in terms of Acc obtained in this study, the results are very close to the results of the study proposed by Kim and Cho [4] even though the C-NSA in this study is not necessary heavy computation load and time just as a deep learning model the C-LSTM in [4] requires heavy computation process. Moreover, the results in this study are very close to the results of the study proposed by Garg *et al.* [49] in terms of FPR.

In the literature, the dataset where network traffic data in the form of time series are presented are few. Many of the global databases, where network traffic data are shared, are presented with a large number of parameter measurements. Accordingly, attacks and abnormal situations in network traffic are usually possible by evaluating all parameters. The performance of the proposed C-NSA approach in detecting abnormalities in network traffic data was also evaluated on the Computer Network Traffic dataset [50]. This dataset contains normal and suspicious network traffic flows from ten different IP hosts on different dates. It has been reported that multiple network flows from five different hosts on four different dates were suspicious for the dataset. Traffic data showing network flows in this dataset is divided into windows consisting of 120 points, as in the C-NSA method proposed in this study. As a result of the experimental studies carried out with the dataset, the malicious network traffic flows from the hosts at IP addresses 1, 3, 4, 5 and 6 using the C-NSA method were determined with an accuracy rate of 97.5%.

5 Conclusion and discussions

The detection of cyber attacks and the prevention are still discussed. Nowadays, it is really important to classify the data and determine whether the data is really harmful or not. Therefore, it is vital to detect and then take action in order to avoid any loss of data of the company and/or person. In this study, C-NSA approach was proposed for the detection of abnormal web traffic on the network and an application software was developed to perform the detection process. For web traffic, the real data in the Yahoo Webscope S5 dataset was used and the data was divided into windows by the sliding window. In the experimental studies, the detection of abnormal traffic data in the web traffic data was provided by monitoring the changes in the number of activated detectors found in the structure of the C-NSA. With the proposed C-NSA method, it was observed that the average performance of detecting the anomalies correctly in real web traffic data in A1 class was 94.30% on average and the performance of the overall classification rate was 98.22% on average. In addition, the results in this study in synthetic web traffic time series in A2 class using C-NSA were achieved as average 96.82% AFR and 98.57% OCA, respectively.

The contributions of this proposed study are as follows. The anomaly traffic in network can be achieved with a proposed method C-NSA with high detection performance. When the studies in the literature are investigated, it is seen that these detection procedures for anomaly detection in network traffic are handled separately and various methods are proposed for each procedure. On the other hand, the proposed method enables performing all procedures with high performance by using C-NSA. In addition, previous studies are improved with the proposed method and high performance has been achieved in the detection of web traffic anomaly. Moreover, in this study, the performance results are similar to the results of previous studies on Yahoo Webscope S5 dataset. The C-NSA can perform extremely fast as it is not necessary heavy computation load and time. In test stage, a detection procedure on a computer specified its configurations in previous sections could be finished by C-NSA in ∼2 s. The proposed approach C-NSA is advantageous in terms of usability in the IDS as compared to complex methods requiring heavy complexity procedures. The fact that web traffic time series in the dataset are obtained from real network data and synthetic data enabled the proposed approach to undergo a detailed test process in accordance with the test conditions.

There are many types of attacks in network infrastructure. In this study, since the data related to each attack type could not be found, the study was conducted only with the data in Yahoo Webscope S5 dataset. This limitation can be said as the missing aspect of our study. In subsequent studies, after obtaining data according to the characteristics of network layers, abnormal traffic can be detected by applying C-NSA. Moreover, the application can be improved by adding more selective features in data classification and application. Besides, abnormal attacks can be detected among various attacks by analysing the attacks. Although IDS can use network traffic data for detection of attacks and anomalies, it is much time consuming [51]. Because the C-NSA is fast and is not required heavy computation load and complexity, it is thought that C-NSA for abnormal traffic can perform the task of web application firewall positioned in front of IDS and even servers. In a more comprehensive study, this proposed method may play an important role in detecting abnormal traffic caused by many attacks.

6 Acknowledgments

The author of this study thanks Yahoo for allowing the use of S5-A labelled anomaly detection dataset [15].

7 References

[1] Papadopoulos, S., Moustakas, K., Drosou, A., *et al.*: 'Border gateway protocol graph: detecting and visualising internet routing anomalies', *IET Inf. Sec.*, 2016, **10**, (3), pp. 125–133

[2] Leiner, B.M., Cerf, V.G., Clark, D.D., *et al.*: 'A brief history of the internet', *ACM SIGCOMM Comput. Commun. Rev.*, 2009, **39**, pp. 22–31

[3] Han, W., Xue, J., Yan, H: 'Detecting anomalous traffic in the controlled network based on cross entropy and support vector machine', *IET Inf. Sec.*, 2019, **13**, (2), pp. 109–116

[4] Kim, T.Y., Cho, S.B: 'Web traffic anomaly detection using C-LSTM neural networks', *Expert Syst. Appl.*, 2018, **106**, pp. 66–76

[5] Symantec Internet Security Threat Report.: Available at https://www.symantec.com/content/dam/symantec/docs/reports/istr-24-executive-summary-en-apj.pdf, accessed 02-Sep-2019

[6] Dandıl, E., Ilhan, K.: 'Anomaly detection in web traffic data using artificial immune algorithms', *Eur. J. Sci. Technol.*, 2019, **October 2019**, pp. 46–56

[7] Sree, T.R., Bhanu, S.M.S.: 'HAP: detection of HTTP flooding attacks in cloud using diffusion map and affinity propagation clustering', *IET Inf. Sec.*, 2018, **13**, (3), pp. 188–200

[8] Mazel, J: 'Unsupervised network anomaly detection', PhD, Networking and Internet Architecture [cs.NI], INSA de Toulouse, 2011

[9] Berger, V.: 'Anomaly detection in user behavior of websites using Hierarchical Temporal Memories: Using Machine Learning to detect unusual behavior from users of a web service to quickly detect possible security hazards', MSc, Computer Science and engineering, Kth Royal Institute Of Technology School Of Computer Science and Communication, 2017

[10] Lai, Y., Chen, Y., Liu, Z., *et al.*: 'On monitoring and predicting mobile network traffic abnormality', *Simul. Modelling Pract. Theory*, 2015, **50**, pp. 176–188

[11] David, J., Thomas, C.: 'Efficient DDoS flood attack detection using dynamic thresholding on flow-based network traffic', *Comput. Secur.*, 2019, **82**, pp. 284–295

[12] Qin, T., Guan, X., Li, W., *et al.*: 'Monitoring abnormal network traffic based on blind source separation approach', *J. Netw. Comput. Appl.*, 2011, **34**, (5), pp. 1732–1742

[13] Xia, H., Fang, B., Roughan, M., *et al.*: 'A basis evolution framework for network traffic anomaly detection', *Comput. Netw.*, 2018, **135**, pp. 15–31

[14] Jing, X., Yan, Z., Jiang, X., *et al.*: 'Network traffic fusion and analysis against DDoS flooding attacks with a novel reversible sketch', *Inf. Fusion*, 2019, **51**, pp. 100–113

[15] Laptev, N., Amizadeh, S.: 'S5-A labeled anomaly detection dataset', Available at https://webscope.sandbox.yahoo.com/catalog.php?datatype=s&did=70, accessed 09-Jun-2019

[16] Potoček, P.: 'Detection of Malicious Network Behaviour in Encrypted Network Traffic', MSc, Faculty of Electrical Engineering Department of Computer Science, Czech Technical University in Prague, 2017

[17] Ganapathy, S., Kulothungan, K., Muthurajkumar, S., *et al.*: 'Intelligent feature selection and classification techniques for intrusion detection in networks: a survey', *EURASIP J. Wirel. Commun. Netw.*, 2013, **271**, pp. 1–16

[18] Agarwal, N., Hussain, S.Z.: 'A closer look at intrusion detection system for web applications', *Sec. Commun. Netw.*, 2018, **2018**, pp. 1–27

[19] Zheng, Y., Liu, Q., Chen, E., *et al.*: 'Time series classification using multi-channels deep convolutional neural networks'. Int. Conf. on Web-Age Information Management, WAIM 2014, Macau, China, June 16–18 2014, pp. 298–310

[20] Münz, G., Li, S., Carle, G.: 'Traffic anomaly detection using k-means clustering'. GI/ITG Workshop MMBnet, Hamburg, Germany, 13/14 September 2007, pp. 13–14

[21] Thill, M., Konen, W., Bäck, T.: 'Online anomaly detection on the webscope S5 dataset: A comparative study'. IEEE Conf. on Evolving and Adaptive Intelligent Systems (EAIS 2017), Ljubljana, Slovenia, 2017, pp. 1–8

[22] Akbal, E., Ergen, B.: 'Kablosuz yerel alan ağlarında yapay bağışıklık sistemi ile saldiri tespiti ve performans analizi', Available at http://www.emo.org.tr/ekler/8947fab05bee9c5_ek.pdf, accessed 02-Sep-2019

[23] Dutt, I., Borah, S., Maitra, I.: 'Intrusion detection system using artificial immune system', *Int. J. Comput. Appl.*, 2016, **144**, (12), pp. 19–22

[24] Kim, J., Bentley, P.J.: 'Towards an artificial immune system for network intrusion detection: an investigation of dynamic clonal selection'. Proc. of the 2002 Congress on Evolutionary Computation, CEC'02, Honolulu, HI, USA, 2002, pp. 1015–1020

[25] Zhang, Y., Lee, W., Huang, Y-A.: 'Intrusion detection techniques for mobile wireless networks', *Wirel. Netw.*, 2003, **9**, pp. 545–556

[26] Das, R.K., Panda, M., Dash, S., *et al.*: 'Application of artificial immune system algorithms in anomaly detection'. Proc. of Progress in Computing, Analytics and Networking, Bhubaneswar, India, 2018, pp. 687–694

[27] Bakhshinejad, N., Hamzeh, A.: 'Parallel-CNN network for malware detection', *IET Inf. Sec.*, 2019, **14**, (2), pp. 210–219

[28] Ganapathy, S., Sethukkarasi, R., Yogesh, P., *et al.*: 'An intelligent temporal pattern classification system using fuzzy temporal rules and particle swarm optimization', *Sadhana*, 2014, **39**, (2), pp. 283–302

[29] Kavin, B. P., Ganapathy, S., Karman, A: 'An intelligent task scheduling approach for cloud using IPSO and A* search algorith'. IEEE Eleventh Int. Conf. on Contemporary Computing (IC3), Noida, India, August 2018, pp. 1–5

[30] Ganapathy, S., Kulothungan, K., Yogesh, P., *et al.*: 'A novel weighted fuzzy C–means clustering based on immune genetic algorithm for intrusion detection', *Procedia Eng.*, 2012, **38**, pp. 1750–1757

[31] Selvi, M., Logambigai, R., Ganapathy, S., *et al.*: 'An intelligent agent and FSO based efficient routing algorithm for wireless sensor network'. IEEE second int. Conf. on recent trends and challenges in computational models (ICRTCCM), Tindivanam, India, February 2017, pp. 100–105

[32] Yahmed, Y.B., Bakar, A.A., Hamdan, A.R., *et al.*: 'Adaptive sliding window algorithm for weather data segmentation', *J. Theor. Appl. Inf. Technol.*, 2015, **80**, p. 322

[33] Atay, Y., Kodaz, H.: 'Optimization of job shop scheduling problems using modified clonal selection algorithm', *Turkish J. Electr. Eng. Comput. Sci.*, 2014, **22**, (6), pp. 1528–1539

[34] Hosseinpour, F., Amoli, P.V., Farahnakian, F., *et al.*: 'Artificial immune system based intrusion detection: innate immunity using an unsupervised learning approach', *Int. J. Digit. Content Technol. Appl.*, 2014, **8**, p.1

[35] Twycross, J.P.: 'Integrated innate and adaptive artificial immune systems applied to process anomaly detection', PhD, University of Nottingham Nottingham, UK, 2007

[36] Timmis, J., Knight, T., de Castro, L.N., *et al.*: 'An overview of artificial immune systems', in Paton, R., Bolouri, H., Holcombe, M., *et al.* (Eds.): '*Computation in cells and tissues*' (Springer, Berlin, Heidelberg, 2004), pp. 51–91

[37] Dasgupta, D.: 'An overview of artificial immune systems and their applications', in, Dasgupta, D. (Ed.): '*Artificial immune systems and their applications*' (Springer, Berlin, Heidelberg, 1993), pp. 3–21

[38] De Castro, L.N., Timmis, J.: '*Artificial immune systems: a new computational intelligence approach*' (Springer, Berlin, Heidelberg, 2002)

[39] Sertkaya, C., Yurtay, N.: 'Artificial immune system based wastewater parameter estimation', *Turkish J. Electr. Eng. Comput. Sci.*, 2018, **26**, (6), pp. 3356–3366

[40] Dandıl, E., Güngör, O.: 'Yapay bağışıklık algoritmaları ile CNC kesici takım aşınmalarındaki değişimin belirlenmesi'. Semp. on Akıllı Sistemlerde Yenilikler ve Uygulamaları (ASYU 2012), Trabzon, Turkey, 2012

[41] Calış, H., Çakır, A., Dandil, E: 'Artificial immunity-based induction motor bearing fault diagnosis', *Turkish J. Electr. Eng. Comput. Sci.*, 2013, **21**, (1), pp. 1–25

[42] Forrest, S., Perelson, A.S., Allen, L., *et al.*: 'Self-nonself discrimination in a compute'. Proc. of 1994 IEEE Computer Society Symp. on Research in Security and Privacy, 1994, pp. 202–212

[43] Ji, Z., Dasgupta, D.: 'Real-valued negative selection algorithm with variable-sized detectors'. Genetic and Evolutionary Computation Conf., Seattle, WA, USA, June 26–30 2004, pp. 287–298

[44] Balachandran, S., Dasgupta, D., Nino, F., *et al.*: 'A framework for evolving multi-shaped detectors in negative selection'. 2007 IEEE Symp. on Foundations of Computational Intelligence, Hawaii, USA, April 2007, pp. 401–408

[45] Jinquan, Z., Xiaojie, L., Tao, L., *et al.*: 'A self-adaptive negative selection algorithm used for anomaly detection', *Prog. Nat. Sci.*, 2009, **19**, pp. 261–266

[46] De Castro, L.N., Von Zuben, F.J.: 'Learning and optimization using the clonal selection principle', *IEEE Trans. Evol. Comput.*, 2002, **6**, pp. 239–251

[47] Ada, G.L., Nossal, S.G.: 'The clonal-selection theory', *Sci. Am.*, 1987, **257**, pp. 62–69

[48] Gao, X.Z., Ovaska, S.J., Wang, X., *et al.*: 'Clonal optimization-based negative selection algorithm with applications in motor fault detection', *Neural Comput. Appl.*, 2009, **18**, pp. 719–729

[49] Garg, S., Kaur, K., Batra, S., *et al.*: 'A multi-stage anomaly detection scheme for augmenting the security in Iot-enabled applications', *Future Gener. Comput. Syst.*, 2019, **104**, pp. 105–118

[50] Computer Network Traffic Dataset: Available at https://www.kaggle.com/crawford/computer-network-traffic, accessed 05-Feb-2020

[51] Sharma, R., Singla, R.K., Guleria, A.: 'A new labeled flow-based DNS dataset for anomaly detection: PUF dataset', *Procedia Comput. Sci.*, 2018, **132**, pp. 1458–1466