International Conference on Cyber Security and Digital Forensics (ICONSEC) 2021

PROCEEDINGS BOOKS

COMMITTEES

Honorary Board

- Prof. Dr. Suat CEBECİ (Yalova University)

Conference Chair

- Prof. Dr. Murat GÖK (Yalova University)

Organizing Committee

- Dr. İrfan KÖSESOY (Yalova University)
- Emre SADIKOĞLU (Yalova University)
- Emine CENGİZ (Yalova University)
- Hasibe CANDAN (Yalova University)
- Fatih BULDUR (Yalova University)

Scientific Committee

- Prof. Dr. Ayhan İSTANBULLU (Balıkesir University)
- Prof. Dr. Ecir Uğur KÜÇÜKSİLLE (Süleyman Demirel University)
- Prof. Dr. Müfit ÇETİN (Yalova University)
- Prof. Dr. Ramazan BAYINDIR (Gazi University)
- Prof. Dr. Naci GENÇ (Yalova University)
- Prof. Dr. Engin AVCI (Fırat University)
- Prof. Dr. Yıldıray YALMAN (Piri Reis University)
- Prof. Dr. Resul DAŞ (Fırat University)
- Prof. Dr. Abdül Halim ZAİM (İstanbul Commerce University)
- Prof. Dr. Ahmet Bedri ÖZER (Fırat University)
- Prof. Dr. Ahmet ZENGİN (Sakarya University)
- Prof. Dr. Atilla ELÇİ (Hasan Kalyoncu University)
- Prof. Dr. Muharrem Tolga SAKALLI (Trakya University)
- Assoc. Prof. Dr. Bilgin METIN (Bogazici University)
- Assoc. Prof. Dr. Abdülkadir TEPECIK (Yalova University)
- Assoc. Prof. Dr. Ahmet KOLTUKSUZ (Yaşar University)
- Assoc. Prof. Dr. Derya AVCI (Fırat University)
- Assoc. Prof. Dr. Sunay TÜRKDOĞAN (Yalova University)
- Assoc. Prof. Dr. Fatih ERTAM (Fırat University)
- Assoc. Prof. Dr. Güzin ULUTAŞ (Karadeniz Technical University)
- Assoc. Prof. Dr. Muhammed Ali AYDIN (İstanbul University)
- Assoc. Prof. Dr. Serdar SOLAK (Kocaeli University)
- Assoc. Prof. Dr. Bünyamin CİYLAN (Gazi University)
- Assoc. Prof. Dr. Ercan BULUŞ (Tekirdağ Namık Kemal University)
- Assoc. Prof. Dr. Sedat AKLEYLEK (Ondokuz Mayıs University)
- Assoc. Prof. Dr. Fatih ÖZKAYNAK (Fırat University)
- Assoc. Prof. Dr. Murat ARICI (Selçuk University)
- Assist. Prof. Dr. Mert ÖZARAR (HAVELSAN Cyber Security Director / Ankara Science University)
- Assist. Prof. Dr. Adem TUNCER (Yalova University)
- Assist. Prof. Dr. Bülent TUĞRUL (Ankara University)
- Assist. Prof. Dr. Burcu DEMIRELLI OKKALIOĞLU (Yalova University)
- Assist. Prof. Dr. Meltem KURT PEHLIVANOĞLU (Kocaeli University)
- Assist. Prof. Dr. Murat AK (Akdeniz University)

- Assist. Prof. Dr. Murat KARAKUŞ (Bayburt University)
- Assist. Prof. Dr. Güneş HARMAN (Yalova University)
- Assist. Prof. Dr. Önder ŞAHINASLAN (Maltepe University)
- Assist. Prof. Dr. Erhan AKBAL (Fırat University)
- Assist. Prof. Dr. Ali DURDU (Social Sciences University of Ankara)
- Assist. Prof. Dr. Murat OKKALIOĞLU (Yalova University)
- Assist. Prof. Dr. Ömer AYDIN (Manisa Celal Bayar University)
- Assist. Prof. Dr. Ömer Özgür BOZKURT (Turkish National Defense University)
- Assist. Prof. Dr. Osman Hilmi KOÇAL (Yalova University)
- Assist. Prof. Dr. Faruk BULUT (Istanbul Rumeli University)
- Assist. Prof. Dr. Süleyman UZUN (Sakarya University of Applied Sciences)
- Assist. Prof. Dr. Şebnem ÖZDEMIR (Beykent University)
- Assist. Prof. Dr. Mustafa COŞAR (Hitit University)
- Assist. Prof. Dr. Tarık YERLIKAYA (Trakya University)
- Assist. Prof. Dr. Yunus ÖZEN (Yalova University)
- Assist. Prof. Dr. Kevser OVAZ AKPINAR (Sakarya University)
- Assist. Prof. Dr. Esra N. YOLAÇAN (Eskişehir Osmangazi University)
- Assist. Prof. Dr. Mustafa Cem KASAPBAŞI (İstanbul Commerce University)
- Assist. Prof. Dr. Fatma BÜYÜKSARAÇOĞLU SAKALLI (Trakya University)
- Assist. Prof. Dr. Atila BOSTAN (Ankara Science University)
- Assist. Prof. Dr. Andaç MESUT (Trakya University)
- Assist. Prof. Dr. Burcu YILMAZEL (Eskişehir Technical University)
- Assist. Prof. Dr. Mehmet Tahir SANDIKKAYA (Istanbul Technical University)
- Assist. Prof. Dr. Alpay DORUK (Bandırma University)
- Assist. Prof. Dr. Özgür Can TURNA (Istanbul University-Cerrahpaşa)
- Assist. Prof. Dr. Mustafa KAYA (Fırat University)
- Dr. Galip Savaş İLGİ (Near East University)
- Dr. Ahmet Ali SÜZEN (Isparta University of Applied Sciences)
- Dr. Mehmet Yavuz YAĞCI (Istanbul University)
- Dr. Remzi GÜRFİDAN (Isparta University of Applied Sciences)
- Dr. Emre Cihan ATEŞ (Gendarmerie and Coast Guard Academy)
- Dr. Ömer ASLAN (Siirt University)
- Dr. Faruk Süleyman BERBER (Süleyman Demirel University)
- Dr. Yunus KORKMAZ (Dicle University)
- Dr. Semih ÇAKIR (Zonguldak Bülent Ecevit University)
- Dr. Kerem GENCER (Karamanoğlu MehmetBey University)
- Dr. Çiğdem BAKIR (Erzincan Binali Yıldırım University)
- Dr. Gülsüm AKKUZU KAYA (Recep Tayyip Erdoğan University)
- Dr. Ahmet KARAKÜÇÜK (Uludağ University)
- Dr. Ömer DURMUŞ (Samsun University)
- Dr. Oğuzhan KENDİRLİ (Düzce University)
- Dr. Duygu Sinanç TERZİ (Amasya University)
- Dr. Mehmet Mehdi KARAKOÇ (Ağrı University)
- Dr. Kerime Dilşad ÇİÇEK (Ayvansaray University)
- Dr. Sultan ZAVRAK (Düzce University)
- Dr. Maad M. MIJWIL (Baghdad College of Economic Sciences University)
- Esra SÖĞÜT (Gazi University)

CONTENTS

International Conference on Cyber Security and Digital Forensics (ICONSEC) 2021

PROCEEDINGS BOOKS

Volume 1

ABSTRACT BOOK

Vulnerabilities of Intrusion Detection and Prevention Systems in Network Traffic Analysis

Mustafa Coşar[1]*

[1]*Hitit University, Çorum, TURKEY*

Abstract

As the connection of people and devices with each other increases day by day, security becomes more important. Standard security measures address the user and attackers. However, this point of view is left helpless with the change in attack methods, the increase in the frequency of attacks and the emergence of new vulnerabilities. In addition, vulnerabilities that may arise from the structures of security systems are also considered important. Some of the weaknesses of IDS and IPS, which are security systems, are slowing down the operation of the system by generating many false positive alarms and breaking the integrity of data packets with fragmentation attacks. In this study, circumvention attacks and features of security systems are mentioned.

Keywords: *Network Traffic Analysis, IDS, IPS, Vulnerability*

* Corresponding author: mustafacosar@gmail.com

Prediction of Ransomware Using Machine Learning Algorithms

Volkan Okur[1]*, Murat Gök[1]

[1]*Yalova University, Yalova, TURKEY*

Abstract

Ransomware is among the main malware today. The detection of ransomware is very important in terms of cyber security. In this study, ransomware types have been estimated by using machine learning methods. In this way, in order to classify ransomware, we first applied the genetic algorithm attribute selection method on the ransomware data set containing 1524 samples and 30,967 attributes and reduced the number of attributes to 7977. Afterwards, we classified them with various learning algorithms. Bagging algorithm gave the best performance with 58.86 % accuracy, 0.55 kappa value and area values under 0.65 precision-precision curve.

Keywords: *Ransomware, Malware, Classification, Bagging, Genetic Algorithm, Attribute Selection*

* Corresponding author: volkanokur@hotmail.com

An Empirical Framework for Fake News Generation

Onur Dura[1]*, Mahmut İmray[1]*, Fatma Gümüş[1]

[1]*National Defence University, İstanbul, TURKEY*

Abstract

Fake news produced to influence social perception, misleading users, or disturbing public peace by spreading the news on patterns and social networks have become an increasingly common problem. The fake news industry uses elements of human and machine learning algorithms. In this study, fake news production with LSTM and BERT models was experimentally investigated. In the continuation of the study, it is aimed to realize an end-to-end system that distorts real news, and to benefit from the system outputs in detecting machine production news in the longer term.

Keywords: *Natural Language Processing, Language Model, Text Generation, LSTM, BERT*

* Corresponding author: 5725dura@harbiyeli.hho.edu.tr

Reinforcement Learning for Cyber Security

Emine Cengiz[1]*, Murat Gök[1]

[1]*Yalova University, Yalova, TURKEY*

Abstract

With the development of today's devices in both software and hardware, development of malicious software and cyber attacks against them are also increasing. Attackers are developing more sophisticated and more complex ways to attack systems every day. Traditional computer algorithms used in cyber security are seem to be insufficient for a solution. Therefore, there is a need for artificial intelligence-based methods. As a matter of fact, Reinforcement Learning (RL), which is a sub-branch of machine learning, is actively used in this field. In this study, we examined the studies in the fields of penetration testing, intrusion detection systems and cyber attacks of RL.

Keywords: *Cyber Attack, Cyber Security, Reinforcement Learning, Penetration Test, IDS*

* Corresponding author: emine.cengiz@yalova.edu.tr

Revealing Vulnerability of Different Machine Learning Models Using Reverse Engineering Method

Emre Sadıkoğlu[1]*, Burcu Demirelli Okkalıoğlu[1], İrfan Kösesoy[2]

[1]*Yalova University, Yalova, TURKEY*
[2]*Kocaeli University, Kocaeli, TURKEY*

Abstract

Nowadays, machine learning-based software is increasing rapidly in every field. These software, developed without considering security, become an obvious target for cybercrime. Generally, rule-based measures such as query restrictions are put in place to keep the server away from adversaries, but the security problem has not been completely overcome. In these systems, the attacker can get output values by sending queries to the classifier to access the information. It explores information about the system by processing the inputs and outputs it obtains without exceeding the query boundary. It then uses the explored information to manipulate the system by attacking the system. In this study, we have revealed the security vulnerabilities of machine learning models, which are created from independent classifying algorithms and do not contain any security measures, by reverse engineering methods on various data sets. Among the manipulated classifiers, the classifier using the SVM method gave the highest class accuracy reduction rate with 16.7% on the Diabetes data set.

Keywords: *Cyber Security, Adversary Attack, Classifier, Security Vulnerability, Machine Learning*

* Corresponding author: emre.sadikoglu@yalova.edu.tr

International Conference on Cyber Security and Digital Forensics (ICONSEC) 2021

PROCEEDINGS BOOKS

Volume 2

FULLTEXT BOOK

Two-step Authentication with the Help of a Novel Biometric

Hidayet TAKÇI[1], Ekin EKİNCİ[2*]

[1] *1Sivas Cumhuriyet University, Faculty of Engineering, Computer Engineering Department, Sivas, TURKEY*

[2] *Sakarya University of Applied Sciences, Faculty of Technology, Computer Engineering Department, Sakarya, TURKEY*

Abstract

In certain cases, one step authentication is not sufficient enough to protect identity. In order to protect the identity there is an urgent need for a secondary authentication mechanism. In this study, a two-step authentication mechanism is proposed. The first step of it is password authentication and the other step is biometric identification mechanism. The novel biometric identification mechanism, discussed in the proposed study, is based on the writing styles. This biometric system includes three feature set: characters, words, and combination of characters and words. The performance of the system is tested on Turkish texts and the performance of feature set composed of combination of characters and words is recorded as 91.90%.

Keywords: *Two-Step Authentication, Biometric Identification, Writing Styles, Classification*

1 Introduction

Recent technological developments provide numbers of benefits as well as a big threat for traditional security systems. In this extend, a lot of methods have been developed to decrease the risk such as passwords, smart cards, pin numbers, open key systems, digital signatures, secure socket layers (SSLs), IP-Sec, secure shell, Kerberos, SSH, and biometric methods [1-3]. The most common security systems are currently password authentication, pin numbers, smart cards and biometric identification systems. However, these systems have their own weaknesses and securities for example the pin and password can be forgotten and hacked as well as smart card authentication systems. One solution to overcome these issues is using key paths producing passwords for only one-use. All improvements in passwords and pins technology, however, have security limitations [4]. Smart card authentication systems are one level more secure and can provide identification, authentication, data storage and application processing. The biggest drawback of smart card is that it can be stolen and copied. The limitations of all these systems shifted the attention to the biometric authentication systems. Biometric authentication systems are based on the physical characteristic of users such as fingerprint, finger vein, retina, palm vein, hand geometry are some of [1]

biological characteristics as well as the behavioural characteristics such as gait style, mouse clicking, keystroke, signature, handwriting and speech [5-7].

In the literature, the advantages and disadvantages of authentication systems have been studied [8]. As an example; Oluwafemi and Feng [9] claimed that, despite of the limitations and weaknesses, the password and pin authentication systems are the more preferred methods by users. In this context, Abbott and Patil [10] reported that although the users' first preferences still are password authentication systems, combined authentication is becoming increasingly important due to security vulnerabilities. Combined systems with passwords and biometric authentication systems would be the best solution to increase the security [11]. In order to decrease the application and hardware cost the most suitable solution for biometric authentications would be behavioural biometrics for example hand-writing, mouse click style.

In the current study, a combined identification system, password and behavioural biometric authentication, was proposed in order to increase

* Corresponding author: ekinekinci@subu.edu.tr

the password security of the online platform. The proposed biometric authentication system is based on users' writing styles that include a summary of lexical, content based and structural characteristics of texts. Hitherto writing style has been used for authorship attribution as well as a biometric authentication system because of its characteristics meeting the demands of biometrics [5]. Current experiments were performed on Turkish text documents and F-measure technique was used in order to test the accuracy and performance of the biometric authentication system. The results showed that combined password application is not only more secure but also more flexible for digital applications.

The rest of the paper is organized as follows: Section 2 summarizes the literature on behavioural biometrics, writing styles and writing styles as a behavioural biometrics. Section 3 details proposed architecture. Section 4 describes experimental study and Section 5 concludes the study.

2 Related works

2.1 Literature on behavioural biometrics

Biometric technologies are performed on biological features such as face, iris, retina, vein patterns, body temperature, or even DNA code and behavioural characteristics such as voice, signature and keystroke dynamics [12, 13]. Biological and behavioural characteristics of users are two main categories of biometric authentication systems. The focus of current study is the behavioural biometric authentication. Therefore, we will only discuss about behavioural characteristics such as voice verification, signature verification, gait recognition, and keystroke dynamics.

Voice verification is one of the most advanced biometric systems is based on voice patterns of a user [14]. Voice based biometric systems consists of three steps: recording the user's voice, identification of voice patterns and classification of voices patterns. The system verifies the individual's identity by comparing the live voice to the stored voice samples.

Signature verification is another common behavioural biometric authentication system [15]. These systems use user's unique signature characteristics such as speed, pen pressure, direction, position and stroke length when an individual inputs the signature [16, 17]. Signature verification is a pattern recognition system that uses signature biometric. Dynamic Time Warping (DTW) and Hidden Markov Model (HMM) algorithms are two methods used to identify the signature characteristics [18, 19]. DTW usually uses a distance measure to compare sequences of integers and different lengths for signature based on distortion [20-23]. Coping with the variance of the signature data HMM is also very successful in verifying signatures [24-26].

Gait biometrics is used to recognize a person from his/her walking and is difficult to change compared to other methods [27]. Gait recognition is realized on time series generated by sensors or sequence of images obtained from video. In recent years, deep learning methods have gained great attention in this area [27-30].

Keystroke dynamics identifies an individual by his/her ubiquitous typing rhythm. A reference template is created and stored in the database by typing the password several times to calculate the average latency between keystrokes. This template is then compared with the currently entered information to verify identity [31]. Alsultan et al. [32] used keystroke dynamics based on timing features such as key press and key release on Arabic tests. Extracted features are the processed in decision trees and ant colony optimisation based support vector machines (SVM). Results showed that SVM was superior than decision trees on the error analysis. Kim et al. [33] realized keystroke dynamics with PIN-based authentication. They applied feature scoring based feature selection method in keystroke dynamics. When they evaluated the classification performance, they found that their methods were 21.8% more successful than the existing ones.

A different method on behavioural biometric is user substitution detection (USD). USD method learns normal behaviour of users on mobile devices and it detects abnormal changes on user behaviour. Mazhelis and Puuronen [34] presented a framework for mobile devices using USD method. Another different behavioural characteristic is handwriting. Dispersion Matcher [35] developed a method used for zip code recognition. Method trains a quadratic discriminant classifier to solve dichotomy problem (whether the two feature vectors belong to the same person). Proposed system was also used on hand geometry and facial recognition problems successfully.

2.2 Writing styles

Historically, writing style is first used to identify authorship attribution of anonym texts. For example; Holmes [36] presented sentence length, word length, word richness of the characters, richness frequencies and dictionary as writing styles. Later, Rudman [37] proposed about 1000 stylistic measurements for authorship attribution. Currently, writing styles have been used in areas such as civil law, criminal law, forensic analysis as well as modern techniques such as information retrieval, machine learning and natural language processing [38, 39]. A lot of research studies have been done in this field. A few important studies in this area are presented below.

Writing styles are the discriminative descriptor features presenting of basic characteristics of texts. Sometimes they are called as style markers. Some researchers used punctuation marks, average word length, average sentence length, non-lexical style markers, sentence boundaries [40, 41] as well as grouping features as writing styles. Peng et al. [42] used features such as n-grams (e.g., bi-grams (groupings of two characters/words), tri-grams (grouping of three characters/words), etc.) as writing styles.

Takçı and Ekinci [43] preferred character based lexical features for authorship attribution of Turkish news. The study emerged from the idea that the character usage frequency of an author is similar in all his/her texts. The feature set composed of 42 features including all letters in Turkish alphabet, punctuation marks and special characters. Identification accuracy of their method is 86.00% in average.

Yavanoğlu [44] applied Artificial NN and Levenberg Marguardt based classifier to identify authorship of Turkish newspaper articles in different genres. To achieve this aim, a total of 41 features from lexical, structural, syntactic, content-specific and idiosyncratic writing styles were used in the study. Politic genre yielded the best accuracy rate (98%).

Varela et al. [45] used syntactical features on literary texts in five languages: Portuguese, Spanish, French, German, and English. In their study, 132 syntactical features, grouped into five categories: (1) morphologic, (2) flexion, (3) syntactic, (4) syntactic auxiliary, and (5) syntactic distances. Among these five languages, identification and verification accuracy of these writing styles for Portuguese are as high as 98.00% and 93.00% respectively.

Ahmed et al. [46] proposed a framework for author attribution of poetries in Arabic. They used a rich set of writing styles including lexical features, character features, structural features, poetry features, syntactic features, semantic features and specific word features. Linear discriminant analysis (LDA), SVM and Naive Bayes was used as classification algorithms. Based on experiments the average best performance was obtained with specific words with accuracy level of 97.95%.

Al-Sarem et al. [47] devised a method for instance-based authorship attribution of Arabic enquires dataset. They used two types features: 397 stylometric features (character, word, syntactic, structural, content, POS, aspect, case, gender, mood, number, grammatical person, state, voice) and 350 distinct words. In addition, they applied AdaBoost and Bagging ensemble classifiers balanced and imbalanced datasets. For balanced dataset Adaboost was the best, for imbalanced dataset Bagging was the best.

2.3 Writing styles as a behavioural biometrics

Recent developments in the computing technologies require new technologies to protect personal data and identity on the Internet. These developments do not only provide a lot of benefits for users but also allow cyber thefts to misuse the personal data. In this aspect, biometric authentication comes to forefront. Biometric authentication systems are widely used, reliable and has gained great popularity in terms of secure authentication. Market size of biometric authentication system was USD 10.74 Billion in 2015 and according to Bhartiya et al., this is expected to increase to USD 32.73 Billion by 2022 [48].

Each biometric category has its own advantages and disadvantages. For example, biological biometrics has some disadvantages in terms of time and money cost. For example, DNA analysis takes a long time as one day. Biological biometric systems are not economically advantageous technologies. Behavioral biometric systems are usually cheaper than biological biometric systems.

In this study, we propose a behavioral biometric authentication system for online environments such as web forums and web content management systems. This biometric authentication system is

based on writing styles of users' online texts. The advantages using proposed biometric can be summarized in three items; (i) there is not any additional cost for data acquisition; (ii) there is no need for an acquisition hardware; (iii) this system strengthens the password verification system.

3 Two-step verification in web sites

Traditionally verification is performed in single step such as password, smart card or biometric method. However, one-step verification may not be enough for the cases requires stronger security. In these cases, the combination of password, smart card or biometric systems would be a better solution. In this study two-step verification has been applied using of password protection as well as a novel biometric. This combined verification system includes password verification as the first step and biometric verification as the second step.

On the other hand, password verification is not secure enough to protect data requires high security protection. Therefore, the current study proposes a two-step verification system employing password and biometric protection together (Figure 1).

Biometric identification systems use biological or behavioural characteristics of users [49]. These systems in general consist of four main modules: sensor module, extraction module, matching module and decision module. Sensor module acquires biometric data from persons' biological or behavioural characteristics. Extraction module extracts most representative features of biometric data. Matching module compares extracted features with stored representative features. Decision module is the classification module whether there is matching or non-matching. The proposed biometric system uses behavioural characteristics that are extracted from news writers. This biometric system supports all biometric system modules described above. Typing the characters is considered as sensor module; the most representative characteristics of the text are considered as extraction module. Biometric verification is considered as matching and decision module and the verification module is shown in Figure 2.

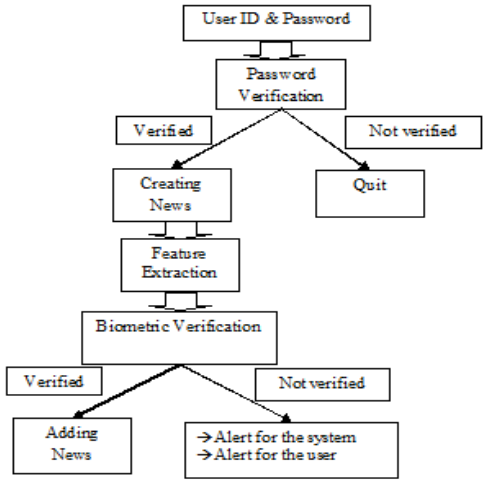


Figure 1. Two-step verification scheme

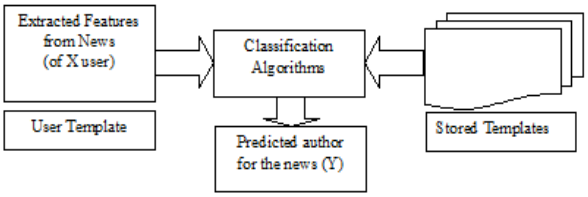


Figure 2. Classification based biometric verification scheme

The proposed biometric has been applied on the classical web content management system. In this system primary control is performed by password, if a user cannot login to system he/she will not have a permission to enter their biometrics, otherwise he/she will be able to enter their biometric identities. In this stage, limited permission such as creating news is given to authenticate users. The user creates the news and posts to the system. After posting, the news has not yet been added to the system. First, the created message is pre-processed via eliminating abusive content, the pre-processed data arrives to the feature extraction module and later message features are extracted. After this operation, biometric comparison is performed. The details are shown in Figure 2.

The purpose of this biometric authentication system is to find the actual author of the news that is created by a person who entered to the system as X. Classification algorithms predict the real author of news. Classification algorithms compare user template (is obtained from the extracted writing styles features from news) and stored templates (which have been obtained from users' reliable texts previously). In this system, there are two types of users: claimed user (X) and predicted user (Y). Claimed user is a person who successfully passes the password verification. Classification algorithms,

based on writing styles of news, detect predicted user. If X=Y then the news pass through the biometric verification. After biometric verification, if the user template is verified (message has been created by the user) message is published in the system; otherwise, there is a serious problem and either the message is suspicious or the password may be stolen or broken. The account owner (legitimate user) is notified of this problem. If the legitimate user approves the suspicious message, the approved message is added to the system. If the legitimate user does not approve this message, the message is deleted and the legitimate user changes his/her password. In the two-step verification, the answers of two questions are important: (1) which features should be used in this system? (2) And which algorithms should be chosen for more performance.

3.1 Feature set

The proposed biometric system uses users' writing styles as biometric characteristics; therefore, writing styles are explained in this section. So far, many features and categories have been proposed for writing styles [50]. In this study, we propose 203 different features in two categories. These categories are character level and word level feature sets. Character level feature set consists of lexical, syntactical features and it includes 66 features. The word level feature set includes 137 features and it contains lexical, content specific and structural features.

3.1.1 Character level features

Character Level Features (CLF) includes byte sized features and their summaries. In this category, there are many lexical, syntactical and structural features. The size of the feature set is 66 and 29 features of the feature set are letters ('a'…'z', 'A'…'Z') and the other 37 features are punctuations marks, space character, several special characters and summary information.

3.1.2 Word level features

Word Level Feature set consists of lexical features (17 words based lexical), structural features (16 words such as greeting message), syntactical features (66 functional words and 4 syntax error based features) and content specific features (34 content specific words).

3.2 Classification algorithms

In the biometric verification template, comparisons between user template and stored templates are performed using classification algorithms. In the literature, many classification algorithms were tested to select the best algorithms by the help of a machine-learning tool [51]. Random Forest (RF), Multilayer Perceptron (MLP), Logistic Regression (LR), Sequential Minimal Optimization (SMO), BayesNet, Naïve Bayes were the most successful algorithms and those were selected for the current study.

Decision tree based algorithms were used in text classification and authorship analysis works [52, 53]. For example, Abbasi and Chen [54] used C4.5 algorithms to analyse English and Arabic group models. They obtained 90.10% for English texts and 71.93% for Arabic texts accuracy rate. Random Forest algorithm that was developed by Breiman is based on decision tree approach. It is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the class's output by individual trees [55]. It is one of the most accurate learning algorithms available. In addition, Random Forest algorithm reported better results than C4.5 algorithms [56, 57]. Therefore, Random Forest was preferred instead of C4.5 for the current study.

Support Vector Machines (SVM) algorithm is one of the most preferred machine learning algorithms in authorship analysis and text categorization. Vapnik [58] introduced the main concepts of (SVM) and Joachims initially used SVM algorithm in text categorization problems [59]. Diederich et al. [60] showed how to use it in authorship analysis. Apart from these, SVM has been used in many authorship attribution studies [46, 61]. One of the SVM related algorithms is Sequential Minimal Optimization (SMO) algorithm which is used for SVM training [62].

In authorship analysis, one of the earliest used methods is statistical method. Francis [63] gave a summary of early statistical approaches used to detect the Federalist Papers' authors. Baayen [64] proposed a linguistic evaluation of statistical models of word frequency. Logistic Regression (LR) [65] is one of the most used statistical classification algorithms. The success of Logistic Regression algorithm has been shown in our experiments.

Neural net approaches were usually used in text categorization problems. Consequently, they are

suitable for authorship analysis. Phani et al. [66] used a multilayer perceptron to attribute authorship to the Bengali blog texts. In our study, Multilayer Perceptron, which is a neural net technique, was used as one of authorship attribution techniques.

4 Experimental study

The purpose of the current experiments is to measure the success of the proposed biometric system. Experiments were conducted with three different feature sets using the software Weka [51].

4.1 Dataset

The dataset has been obtained from a web content management system (WCMS) [67] of which has the users who can post news to the system. This corpus contains 289 news which are posted by 10 users between 2006 March and 2008 April. The length of these texts varies between 159 and 994 words. The reliability of the posted messages was guaranteed in order to provide a robust dataset for current study.

Table 1. User details for data set.

Users	Description
User 1	User 1 writes on similar issues. He uses the correct spelling rules.
User 2	User 1 writes long texts. He uses detailed descriptions and special names in his texts.
User 3	User 3 uses small sentences and he writes long texts. He usually uses the correct spelling rules.
User 4	User 4 writes small texts. He doesn't use the spelling rules.
User 5	User 5 writes literary texts. He sometimes uses poetry samples.
User 6	User 6 writes on different topics. She uses specific punctuation marks.
User 7	User 7 texts' have spelling errors. He writes small texts in a few topics.
User 8	User 8 writes literary texts. He uses the correct spelling rules.
User 9	User 9 writes different topics. He writes average sized texts. He complies spelling rules.
User 10	User 10 writes in a single topic. He writes long texts. He sometimes complies spelling rules.

Some simple pre-processing operations were conducted during importing news from web site to the corpus. For example, quotations were deleted from news texts to obtain reliable texts. Later, each news text has been transformed into writing style-based vectors. These vectors were used in classifier training and testing. After this transformation, writing styles were ready for classification algorithms.

4.2 Experimental design

In this study two-step verification mechanism was conducted and classification algorithms such as Random Forest (RF), Multilayer Perceptron (MLP), Logistic Regression (LR), Sequential Minimal Optimization (SMO), BayesNet, Naïve Bayes performed the second step of this mechanism. These classifiers are trained by users' reliable texts to verify news posted by users.

Three different experiments were conducted based on three different feature sets on this data set to measure the second step verification. In each experiment; RF, C-SVC, LR, MLP, SMO, BayesNet and Naïve Bayes algorithms were performed. In the first run, character level writing styles were used. In the second run, word level writing styles were used and combined (character level + word level features) feature set was used in the third run. At the end of each experiment of learning algorithms were evaluated by 10-fold cross-validation method.

Table 2. Confusion matrix for model evaluation.

	Real Author	Somebody Else
Real Author	True Positive (TP)	False Negative (FN)
Somebody Else	False Positive (FP)	True Negative (TN)

In these experiments; accuracy, recall and precision factors were used as performance metrics. These are most frequently used factors to measure the prediction of authorship performance analysis. Table 4 provides the detailed information about these measures. The first performance measure we used in the experiments is accuracy. Accuracy can be expressed as follows.

$$accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad (1)$$

4.3 Accuracy based results

In the experiments, accuracy matrixes are generated from the classification accuracies for each feature set. All values are shown in Table 3.

Table 3. Classification algorithms and accuracy rates (%) for feature sets.

Classification Algorithm	Character Level Features	Word Level Features	Mixed features
Random Forest	63.06	56.30	64.41
Multilayer Perceptron	80.18	77.93	89.19

Logistic Regression	72.07	82.88	91.90
SMO	77.07	79.23	89.19
BayesNet	67.11	66.21	77.93
Naive Bayes	76.58	60.81	68.02

The highest accuracy rate for the character level feature set is 80.18%. This value has been obtained by MLP algorithm. Character level feature set contains lexical, structural and syntactical features. These features are informative features in authorship analysis. Word level feature set contains syntactical, content specific, lexical features. In addition, this feature set also contains structural features about the special content. The most valuable style markers of word level feature set are lexical and content specific features. Observed results show that the best word level features are structural features and lexical features. The best classification result belongs to Simple Logistic algorithm with 82.88% accuracy rate. In addition, the word-based features are highly author and language dependent.

Third feature set is a combination of first feature set and second feature set. This feature set is addressed as combined or mixed feature set. Combining feature sets improved the classification success. Simple Logistic algorithm gave the best results for combined features with 91.90% accuracy rate. Among all writing style categories, combined level ones seem to carry more information.

4.4 F-measure based results

Accuracy metric sometimes may not be enough for reliable evaluation; therefore, we use F-measure metrics. F-measure metrics is obtained from precision and recall values.

$$Precision(p) = \frac{TP}{TP + FP} \quad (2)$$

$$Recall(r) = \frac{TP}{TP + FN} \quad (3)$$

$$F - measure = \frac{2 \times p \times r}{p + r} \quad (4)$$

F-measure metric was used for feature sets and classification algorithms. Firstly, F measure curves were obtained for character level, word level and combined feature set. This curve was shown in Figure 3.

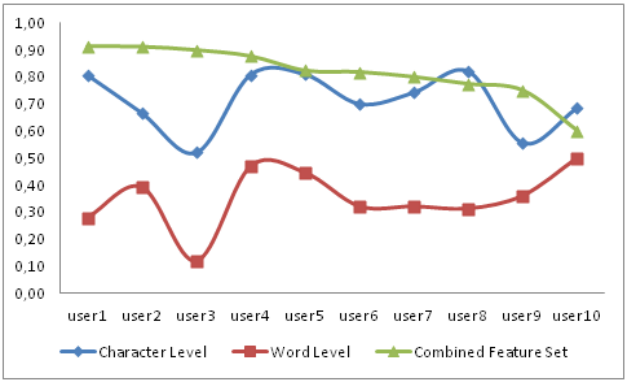


Figure 3. F-measure curves for each writing style categories

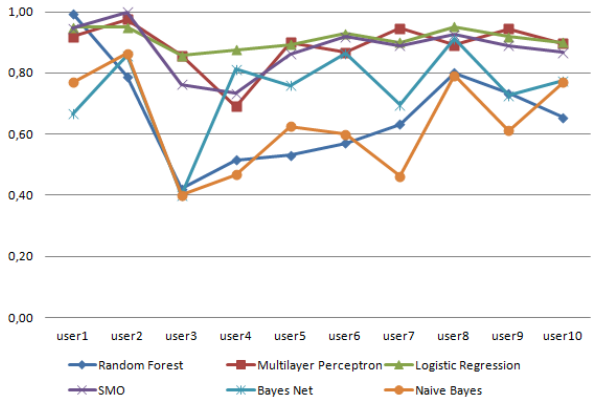


Figure 4. F-measure curves for the classification algorithms

Figure 4 presents F-measure curves for all classification algorithms by using combined feature set. As shown in Figure 4, the best classifiers are Logistic Regression, Multilayer Perceptron and SMO. The worst classifier is Naïve Bayes and Random Forest. In fact, Naïve Bayes and Random Forest algorithms are successful algorithms but the results are bad for this data. This figure also gives some information to us about users. For example, the classification accuracy for User 2 is higher but for User 3 is lower. There are some reasons of this situation: User 2 writes longer texts but User 3 writes smaller texts and longer texts contain the more writing styles. User 2 writes in several special issues (fewer topics) but User 3 writes in many areas. User 2 uses technical terms but User 3 doesn't technical terms. User 2 more educated and he has some characteristic words but User 3 uses generally

spoken words. User 2 complies spelling rules but User 3 doesn't comply.

5 Conclusion

A web content management system (WCMS) is a secure software system that provides website authoring, collaboration, and administration tools designed to allow users. A robust WCMS offers some opportunities to create and manage news documents. However, the challenge is to protect the identity of users on WCMS to prevent unwanted and inappropriate publications. The solution for this problem would be a secondary protection in addition to current password protection systems.

In this study, a novel biometric was proposed as a two-step online identity protection. This biometric authentication system is based on users' writing behaviors on the computer. The feature extraction tool would provide a robust security system to detect actual users and eliminate fraud.

The proposed method was evaluated with three different feature sets and several classification algorithms on Turkish web content management site data with an accuracy rate of 91.90%. Combining feature sets improved the classification success. Although size of word level feature set is bigger than size of the character level feature set, classification results are lower than character level feature set. The reason of this is the size of the whole word level features is very big and here only a part of the possible features was used in our proposed method. The limitation of the current system is the limited number of feature sets. Increasing the number of feature sets would allow us to develop more robust behaviour-based biometric systems.

References

[1] Duncan R, "An Overview of Different Authentication Methods and Protocols", SANS Institute, Swansea, UK, 2001.

[2] de Fuentes JM, Hernandez-Encinas L, Ribagorda A. *Security Protocols for Networks and Internet: A Global Vision*. Editors: Daimi K. Computer and Network Security Essentials, 135–151, Springer Cham, 2007, 2007.

[3] Bhootwala JP, Bhathawala PH. "Graphical Password Authentication – Survey". *Global Journal for Research Analysis*, 9(2), 31-35, 2020.

[4] Yashaswini KP, Rani N, Jadhav V, Ajina A. "Abopi: Authentication by Bringing Own Picture/Image". *2nd IEEE International Conference on Recent Trends in Electronics Information & Communication Technology (RTEICT)*, Bangalore, India, 19–20 May 2017.

[5] Delac K, Grgic M. "A survey of biometric recognition methods". 46th International Symposium Electronics in Marine, ELMAR-2004, Zadar, Croatia, 18 June 2014.

[6] Buciu I, Gacsadi, A. "Biometrics systems and technologies: A survey". *International Journal of Computers Communications & Control*, 11(3), 315-330, 2016.

[7] Dargan S, Kumar M. "A comprehensive survey on the biometric recognition systems based on physiological and behavioral modalities". *Expert Systems with Applications*, 143, 1-27, 2020.

[8] O'Gorman L. "Comparing Passwords, Tokens, and Biometrics for User Authentication". *Proceedings of the IEEE*, 91(12), 2019-2040, 2003.

[9] Oluwafemi AJ, Feng JH. *Usability and Security: A Case Study of Emergency Communication System Authentication*. Editors: Stephanidis C. HCI International 2019 - Posters. HCII 2019. Communications in Computer and Information Science, 205-210, Springer Cham, 2019.

[10] Abbott J, Patil S. "How Mandatory Second Factor Affects the Authentication User Experience", 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 25-30 April 2020.

[11] Deane F, Barrelle K, Henderson R, Mahar D. "Perceived acceptability of biometric security systems". *Computers & Security*, 14(3), 225-231, 1995.

[12] Noh HW, Ahn CG, Kong HJ, Sim JY. "Ratiometric Impedance Sensing of Fingers for Robust Identity Authentication". *Scientific Reports*, 9(13566), 1-12, 2019.

[13] Buckley O, Nurse JRC. "The language of biometrics: Analysing public perceptions". Journal of Information Security and Applications, 47, 112-119, 2019.

[14] Aizat K, Mohamed O, Orken M, Ainur A, Zhumazhanov Z. "Identification and authentication of user voice using DNN features and i-vector". *Cogent Engineering*, 2020, 7(1751557), 1-21, 2020.

[15] Dey S, Dutta A, Toledo JI, Ghosh SK, Llados J, Pal U. "SigNet: Convolutional Siamese Network for Writer Independent Offline Signature Verification". *arXiv:1707.02131*, 2017.

[16] Yahyatabar ME, Ghasemi J. "Online signature verification using double-stage feature extraction modelled by dynamic feature stability experiment". *IET Biometrics*, 6(6), 393-401, 2017.

[17] Saleem M, Kovari B. "Preprocessing Approaches in Dtw Based Online Signature Verification". Pollack Periodica, 15(1), 148-157, 2020.

[18] Parmar M, Puranik N, Joshi D, Malpani S, Thakare B. "State of Art Survey Signature Verification Techniques 2019". *Asian Journal for Convergence in Technology*, 5(3), 91-96, 2019.

[19] Hu H, Zheng J, Zhan E, Tang J. "Online Signature Verification Based on a Single Template via Elastic Curve Matching". *Sensors*, 19(4858), 1-23, 2019.

[20] Munich ME, Perona P. "Continuous Dynamic Time Warping for translation invariant curve alignment with applications to signature verification". *7th International Conference on Computer Vision*, Korfu, Greece, 20-27 September 1999.

[21] Kholmatov A, Yanikoglu B. "Identity authentication using improved online signature verification method". *Pattern Recognition Letters*, 26(15), 2400–2408, 2015.

[22] Tolosana R, Vera-Rodriguez R, Fierrez J, Ortega-Garcia J. "DeepSign: Deep On-Line Signature Verification". *arXiv:2002.10119*, 2020.

[23] Qiao Y, Wang X, Xu C. "Learning Mahalanobis Distance for DTW based Online Signature Verification". *2011 IEEE International Conference on Information and Automation*, Shenzhen, China, 6-8 June 2011.

[24] Wada N, Hangai S. "HMM Based Signature Identification System Robust to Changes of Signatures with Time". *2007 IEEE Workshop on Automatic Identification Advanced Technologies*, Alghero, Italy, 7-8 June 2007.

[25] Chavan M, Singh RR, Bharadi VA. "Online Signature Verification using Hybrid Wavelet Transform with Hidden Markov Model". *2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA)*, Pune, India, 17-18 August 2017.

[26] Kar AK, Chandra SK, Bajpai MK. "Parallel Gpu Based Offline Signature Verification Model". *2019 IEEE 16th India Council International Conference (INDICON)*, Rajkot, India, 13-15 December 2019.

[27] Zou Q, Wang Y, Wang Q, Zhao Y, Li Q. "Deep Learning-Based Gait Recognition Using Smartphones in the Wild". *IEEE Transactions on Information Forensics and Security*, 15, 3197-3212, 2020.

[28] Davarzani S, Saucier D, Peranich P, Carroll V, Turner A, et al. "Closing the Wearable Gap—Part VI: Human Gait Recognition Using Deep Learning Methodologies". *Electronics*, 9(796), 1-17, 2020.

[29] Terrier P. "Gait Recognition via Deep Learning of the Center-of-Pressure Trajectory". *Applied Sciences*, 10(774), 1-20, 2020.

[30] Wang X, Zhang J. "Gait feature extraction and gait classification using two-branch CNN". *Multimedia Tools and Applications*, 79, 2917–2930, 2020.

[31] Lu X, Zhang S, Hui P, Lio P. "Continuous authentication by free-text keystroke based on CNN and RNN". *Computers & Security*, 96, 10186, 2020.

[32] Alsultan A, Warwick K, Wei H. "Free-text keystroke dynamics authentication for Arabic language". *IET Biometrics*. 5(3), 164-169, 2016.

[33] Kim DI, Lee S, Shin JS. "A New Feature Scoring Method in Keystroke Dynamics-Based User Authentications". *IEEE Access*, 8, 27901-27903, 2020.

[34] Mazhelis O, Puuronen S. "A framework for behavior-based detection of user substitution in a mobile context". *Computers & Security*, 26, 154-176, 2007.

[35] Fàbregas J, Faundez-Zanuy M. "Biometric dispersion matcher". *Pattern Recognition*, 41, 3412-3426, 2008.

[36] Holmes DI. "The evolution of stylometry in humanities scholarship". *Literary and Linguistic Computing*, 13(3), 111-117, 1998.

[37] Rudman J. "The state of authorship attribution studies: Some problems and solutions". *Computers in Human Behavior*, 31, 351-365, 1998.

[38] Benzebouchi NE, Azizi N, Hammami NE, Schwab D, Khelaifia MCE, Aldwairi M. "Authors' Writing Styles Based Authorship Identification System Using the Text Representation Vector". *16th International Multi-Conference on Systems, Signals & Devices (SSD'19)*, Istanbul, Turkey, 21-24 March 2019.

[39] Brocardo ML, Traore I, Woungang I. *Continuous Authentication Using Writing Style*. Editors: Obaidat MS et al. Biometric-Based Physical and Cybersecurity Systems, 211-232, Springer Nature Switzerland, 2019.

[40] Ekinci E, Takçı H. "Using authorship analysis techniques in forensic analysis of electronic mails". *20th Signal Processing and Communications Applications Conference (SIU)*, Muğla, Turkey, 18-20 April 2012.

[41] Verma G, Srinivasan BV, "A Lexical Syntactic and Semantic Perspective for Understanding Style in Text". *arXiv: 1909.08349v1*, 2019.

[42] Siagian AHAM, Aritsugi M. "Robustness of Word and Character N-gram Combinations in Detecting Deceptive and Truthful Opinions". *ACM Journal of Data and Information Quality*, 12(1), 1-24, 2020.

[43] Takçı H, Ekinci E. "Character Level Authorship Attribution for Turkish Text Documents". *The Online Journal of Science and Technology*, 2(3), 12-16, 2012.

[44] Yavanoğlu Ö. "Intelligent Authorship Identification with using Turkish Newspapers Metadata". 2016 *IEEE International Conference on Big Data*, Washington, DC, USA, 5-8 December 2016.

[45] Varela PJ, Albonico M, Justino EJR, Bortolozzi F. "A computational approach to authorship attribution in multiple languages". *2018 International Joint Conference on Neural Networks*, Rio de Janeiro, Brazil, 8-13 July 2018.

[46] Ahmed AF, Mohamed R, Mostafa B. "Arabic Poetry Authorship Attribution using Machine Learning Techniques". *Journal of Computational Science*, 15(7), 1012-1021, 2019.

[47] Al-Sarem M, Saeed F, Alsaeedi A, Boulilia W, Al-Hadhrami T, "Ensemble Methods for Instance-Based Arabic Language Authorship Attribution". *IEEE Access*, 8, 17331-17345, 2020.

[48] Bhartiya N, Jangid N, Jannu S. "Biometric Authentication Systems: Security Concerns and Solutions". *3rd International Conference for Convergence in Technology*, Pune, India, 6-8 April 2018.

[49] Gomez-Barrero M, Galbally J, Rathgeb C, Busch C. "General Framework to Evaluate Unlinkability in Biometric Template Protection Systems". *IEEE Transactions on Information Forensics and Security*, 13(6), 1406-1420, 2018.

[50] Al-Khatib MA, Al-qaoud JK. "Authorship verification of opinion articles in online newspapers using the idiolect of author: a comparative study", Information, Communication & Society, 1-19, 2020.

[51] Weka. "Weka - Machine Learning Software in Java". http://sourceforge.net/projects/weka/, (10.04.2020).

[52] Bahassine S, Madani A, Al-Sarem M, Kissi M. "Feature selection using an improved Chi-square for Arabic text classification", *Journal of King Saud University- Science*, 32(2), 225-231, 2020.

[53] Srinivasan L, Nalini C. "An improved framework for authorship identification in online messages". *Cluster Computing*, 22, 12101–12110, 2019.

[54] Abbasi A, Chen H. "Applying authorship Analysis to Extremist-Group Web Forum Messages". *IEEE Intelligent Systems*, 20, 67-75, 2005.

[55] Breiman L. "Random Forests". *Machine Learning*, 45(1), 5-32, 2001.

[56] Brown I, Mues C. "An experimental comparison of classification algorithms for imbalanced credit scoring data sets". *Expert Systems with Applications*, 39(3), 3446-3453, 2012.

[57] Baron G. *Comparison of Cross-Validation and Test Sets Approaches to Evaluation of Classifiers in Authorship Attribution Domain.* Editors: Czachórski T, Gelenbe E, Grochla K, Lent R. Communications in Computer and Information Science, 81-89, Springer, Cham, 2016.

[58] Vapnik V, *The Nature of Statistical Learning Theory*. 2nd ed. New York, USA, Springer-Verlag, 2000.

[59] Joachims T. *Text categorization with support vector machines: Learning with many relevant features*. Editors: Nédellec C, Rouveirol C. Lecture Notes in Computer Science, 137-142, Springer, Berlin, Heidelberg, 1998.

[60] Diederich J, Kindermann J, Leopold E, Paass G. "Authorship Attribution with Support Vector Machines". *Applied Intelligence*, 19, 109-123, 2003.

[61] Nirkhi S. *Evaluation of Classifiers for Detection of Authorship Attribution*. Editors: Verma N, Ghosh A. Computational Intelligence: Theories, Applications and Future Directions - Volume I. Advances in Intelligent Systems and Computing, 227-236, Springer, Singapore, 2018.

[62] Platt JC. Using analytic QP and sparseness to speed training of support vector machines. *Advances in neural information processing systems*, 557-563, 1999.

[63] Francis IS. *An Exposition of a Statistical Approach to the Federalist Dispute*. Editors: Leed J. The Computer and Literary Style, 38–79, Kent, Ohio: Kent State University Press, 1966.

[64] Baayen RH. "Statistical Models for Word Frequency Distributions: A Linguistic Evaluation". *Computers in Human Behavior*, 26, 347–363, 1992.

[65] Le Cessie S, Van Houwelingen JC. "Ridge Estimators in Logistic Regression". Journal of the Royal Statistical Society: Series C, 41(1), 191-201, 1992.

[66] Phani S, Lahiri S, Biswas A. "A machine learning approach for authorship attribution for Bengali blogs". 2016 International Conference on Asian Language Processing (IALP), Tainan, Taiwan, 21-23 November 2016.

[67] Sucati Haber. www.sucati.com, (10.02.2011).

A Model for Header Compression Context Transfer in Cellular IP

Rouba Omar Alahmad Alosman[1*], İsmail Hakkı Cedimoğlu[1]

[1]*Sakarya University, Sakarya, TURKEY*

Abstract

Providing uninterrupted Internet connectivity to mobile users has become a serious challenge in recent as the increasing number of users requires more bandwidth especially in the next generation networks. Time is an equally important factor while providing these services to the mobile users. In mobile computing IP based technologies are extremely important. Mobile IP deals with the management of the IP when the macro mobility is concerned. Cellular networks are used as an access method to the Internet and the other IP-based networks. Cellular IP has been suggested and suited for the micro mobility of the users. Context transfer in Cellular IP protocol incurs an overhead, which often is intolerable as the micro mobility is in concern. The basic idea of the proposed model depends on attaching the context established during Header Compression (HC) process with an IP signaling message and transferring it from base station to another before the handoff operation takes place. Since the proposed model discusses the transfer of Header Compression context, it also shows the importance of Header Compression in reducing the consumed bandwidth through NS-2 simulation experiment. The result reveals the efficacy of the model.

Keywords: *Macro Mobility, Micro Mobility, Mobile IP, IP Header, IP Context, Handoff*

1 Introduction

Mobile IP (MIP) is a protocol that gives any mobile user the ability to roam beyond its network. Mobile IP can be seen as a protocol that provides seamless macro mobility solutions [1]. The phrase Mobile IP can be understood either as the mobilized Internet with all different technologies and operations or it can refer to the MIP protocol defined by IETF (Internet Engineering Task Force). The standard Mobile IP term consists of three entities: Mobile Node (MN), Home Agent (HA) and Foreign Agent (FA). Mobile IP is a very good solution for the mobility problem at the macro level of the mobility where MIP maintains two addresses: home address, Mobile IP is a good solution for the macro level of and foreign address, but the user is considered to be having a single address [3].

Though users' mobility, it is not suitable for the users that move within one network with the micro-level of mobility. The solution that supports local mobility and efficiently inter-works with Mobile IP to provide wide area mobility support is Cellular IP. Cellular IP (CIP) consists of the Base Station (BS), Gateway (GW) and the Mobile Node (MN). In Cellular IP each and every node maintains two kinds of caches: one is used for location management and the other is used for routing management. Cellular IP ca[1]n distinguish between active nodes and the

nodes that don't send or receive any data. It maintains the position of the "idle" mobile in paging cache [4].

Benefits of Cellular IP can be considered as its ability to accommodate a large number of users attached to the network without overloading the location management system. Also distributed location management and routing algorithms tend to be simple in Cellular IP networks. Low cost implementation of Internet host mobility requiring no new packet format or any new operations other than those exist in IP based networks are also its features [4].

Many services can be given to the mobile user when he roams between the networks under MIP protocol or roams from one cell to another under CIP protocol. The mentioned user should get services from the network or sub-network to which he moves away from his own network. To avoid the re-establishment of these services and give them to the user, a Context Transfer protocol is used. It is used so in order to save the time and bandwidth of the

* Corresponding author: rubaalosman@hotmail.com

network. One of the most important services to be provided to the user is Header Compression which is the technique of reducing the size of the IP header at the sender side (compressor) in such a way so that it is possible to reconstruct the compressed header at the receiver end (de-compressor). Header Compression can save the time and the bandwidth of the network and so is an important activity. Header Compression technique was widely adopted and well-studied, many algorithms have been proposed for developing Header Compression as in [5]. Many models have been proposed for Header Compression Context as in [6], where a simulation experiment using OPNET simulator was performed. Their experiment has focused over the effect of context transfer on packet size and time consumed in mobile IP network (macro level of mobility). In [13] two schemes of Context Transfer have been proposed: Context Transfer extension to mobility management protocols and standalone Context Transfer module. In [10] a Context Transfer solution was proposed for transferring state information for candidate service stored at the micro mobility domain gateway to the mobile host's new base station once handoff takes place. However, our proposal has suggested transferring of Context of Header compression in Cellular IP networks (micro level of mobility).

2 Mobility Management Protocols

Mobile IP is the protocol that is most suitable for macro level of mobility and consists of

- Mobile Node which is a user with computing device that changes its point of attachment from one sub-network to another.
- Home Agent is a router found in user's home network and is responsible of providing all services to the users in its network.
- Foreign Agent is found in the visited network and is responsible of providing routing services to the user registered in this network

2.1 Operations in mobile IP

The operations of Mobile IP are elaborated as follows.

2.1.1 Discovery

In discovery operation, ICMP (Internet Control Messages Protocol) messages are used. The agent of a network broadcasts ICMP messages, periodically, as advertisements telling the hosts of other networks that it can serve them in its network. When the Mobile Node receives these messages, it starts comparing its IP address with the IP address of the agent from which it received the ICMP message. The comparison operation is done between the network portions in each IP address. If the address were equal then the Mobile Host will know that it is in its home network. Otherwise, it will discover that it is in a foreign network [1, 5].

2.1.2 Registration

The Foreign Agent, after completion of discovery procedure, will grant a care-of-address to the Mobile Node. The latter will register this care-of-address with the home agent. Registration procedure at time t, as depicted in figure 1.a, takes the following steps.

The Mobile Node sends a registration request to the Foreign Agent. When the Foreign Agent receives a registration request it relays this request to the Home Agent. The Home Agent either accepts or rejects the request. The Foreign Agent relays the reply to the Mobile Node [1, 5]. It may happen, sometimes, that the Mobile Node moves to another network without Foreign Agent or the Foreign Agent is busy at the time. Therefore, the Mobile Node acts as a Foreign Agent and directly registers its care-of-address, called collocated care-of-address.

2.1.3 Tunneling

In this operation the Home Agent adds a new IP address to the packet sent to the Mobile Node. The total packet consists of two headers besides the payload which exists originally in the basic packet. This is the encapsulation operation after which a tunnel will be opened between the Home Agent and the Foreign Agent and the encapsulated packet will be sent through it as evident from figure 2 [1].

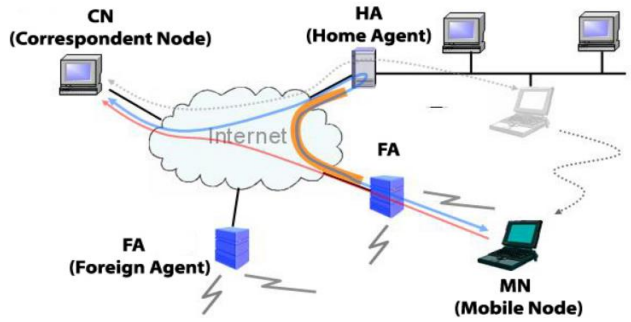


Fig. 2. Tunneling operations in Mobile IP

2.2 Cellular IP

To manage the mobility of users from one cell to another in one network, Cellular IP is used. This protocol gives an efficient location management taking into consideration the passive connectivity, paging aspect, and the fast handoff control. The difference of Cellular IP from Mobile IP is that Cellular IP protocol implements the mentioned principles from IP point of view (IP paradigm). There are some good features for this protocol such as the minimal need of resources, simplicity in design and minimal use of signaling which is very good in sense of not overloading the network [1].

2.2.1 Cellular IP components and operations

As Cellular IP protocol inherits most of the cellular system's features, it is evident that Cellular IP model will be similar to the Cellular networks. Cellular IP Model consists of the following components.

Base Station (BS): which works as a wireless access point. If a Mobile Node is willing to move from one location (cell) to another it has to attach itself with a new access point (new Base Station) and it will be responsible of providing all services to this Mobile Node. It will also make routing of all packets coming to this Mobile Host until they (data packets) reach their destination.

Gateway: Through the gateway, the mobile hosts connect to the internet and also, all mobile hosts are able to communicate with the "Corresponding Nodes" in another network.

Mobile Node: This entity in Cellular IP network moves from one cell under control of one Base Station, to another cell where it has to register with the new Base Station under which it is willing to move through an operation called handoff. Figure 3 shows the basic elements of Cellular IP networks [6].

Operations that take place in Cellular IP networks are as follows.

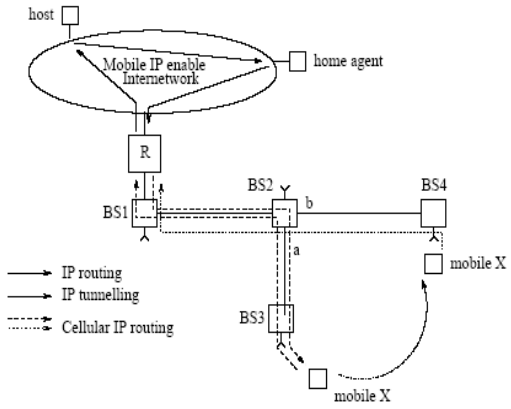


Fig.3 . Cellular IP Model

2.2.2 Handoff

Handoff operation in Cellular IP network refers to the movement of Mobile Node from one geographical area called cell under the control of one Base Station to another cell (Base Station). This operation depends on the strength of the signal coming to the Mobile Node and according to this strength the handoff is decided. There are two types of handoff; hard handoff and soft handoff.

Hard handoff is one in which the Mobile Node measures the strength of the signal coming to it from another Base Station and if it discovers that a signal coming from the new Base Station is stronger than the old one, it connects directly to the new Base Station and operates under the control of the new Base Station.

Soft handoff makes Mobile Node send a soft packet, as a request, to the new Base Station to which it is approaching. Soft packet is an IP packet of a special structure sent from the Mobile Node to the Base Station before performing soft handoff; therefore this special packet will carry the request to the new Base Station and some additional information as discussed in section 4. After sending the soft packet, the Mobile Node returns back at the same time to the old Base Station to continue the session that was taking place between it and the old Base Station. Using the soft packet, the routing cache mapping will be changed in order to establish a new route related to this Mobile Node. At the time of new route establishment, the Mobile Node will be connected to the old Base Station and continue the old session with the old Base Station. This procedure continues

for a soft time delay. After this time delay, the Mobile Node will be connected to the new cell and begins a new session with the new Base Station [8].

3 Header Compression

Header Compression is an important operation and is necessary in Mobile IP networks and Cellular IP networks as well. The importance of this operation yields from the fact that it can improve Quality of Service (QoS) provided to the users in saving bandwidth. Therefore it is important to brief some of the benefits of Header Compression and go through its procedure.

3.1 Header compression and link efficiency

Efficiency of Header Compression is observed by studying IPv4 which is total 40 bytes in size. Of which:
IPv4 size=20 bytes.
User Datagram Protocol (UDP) size=8 bytes.
Real Time Transport Protocol (RTP) size =12 bytes.
It has been observed that by using Header Compression we can decrease the size from 40 Bytes to 2-4 bytes as shown in figure 4 [2].

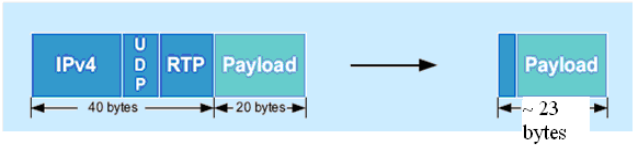


Fig 4. Packet length reduction using Header Compression

Header Compression can improve the efficiency of transmission link by improving response time due to smaller packet size that we get by Header Compression. Also it results in reduction in the probability of packet loss. Efficiency improvement of link can be observed by using Header Compression where Header Compression will decrease packet header overhead, reduces packet loss and improve response time of the network. The benefits of the Header Compression results in efficient link [9].

3.2 Header Compression process

The main principle of Header Compression (HC) is based on the fact that there is a big amount of redundant information in the headers of every packet to be sent. Most of the header values remain same while the session, that contains these packets, is going on. These sessions are classified into two types: non-TCP sessions in which, most of the fields are constant and TCP sessions in which there are several constant fields while the others change in a predictable way. Header fields are classified into four types:

STATIC fields: these fields don't change.
RANDOM fields: these fields are not constant so they can be changed.
DELTA fields: these fields change in a small value called delta value; this value can be sent instead of sending the whole field.
INFERRED fields: the values of these fields can be inferred from other values.
Of these four fields, STATIC field and DELTA field are constant and will form the redundant data that we send every time. We can compress each packet before sending (in the compressor side) by not including STATIC and DELTA fields.
Compression process starts by the establishment of Header Compression (HC) Context at both the compressor and the de-compressor. At the compressor side, the compressor examines the packet headers, copies the values of these headers to establish the HC context. Finally it assigns a Context Identifier CID to the established context and then sends the uncompressed packet with CID to the de-compressor in order to establish the context at its side. Upon receiving a packet with CID, the de-compressor starts to build the HC context using packet header fields. The HC context will consist of STATIC fields and DELTA fields. Once the HC context is established, packet headers can be compressed, as we have mentioned previously, by not including STATIC fields and by using fewer bits to store delta values. Establishment of the HC context must be done through an enhanced scheme to ensure the establishment of the HC context correctly. One of the suggested schemes to establish HC context is three state HC context as shown in figure 5 [10].

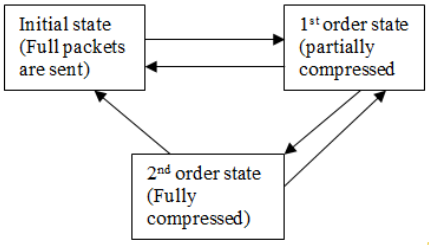


Fig.5. Three States HC Context

In the partially compressed packet, the packet header consists of CID, RANDOM fields, DELTA fields and INFERRED fields; therefore HC context consists of STATIC fields. In the fully compressed packet, the packet header consists of CID, RANDOM fields, INFERRED fields and delta values; so STATIC fields and DELTA fields form the HC context [10, 11].

3.3 Context Transfer

In order to be efficient, Cellular IP network must provide many services to its users such as Authentication, Accounting, Authorization (AAA) and Header Compression, etc. Each one of these services has context. If these contexts are to be established each and every time, there will be a great consumption of the user's time and the network's time as well; therefore Context Transfer is suggested as a solution for this problem.
Context differs for different services and before transferring it, context must be established at the sender and receiver sides. After establishment of the context it has to be transferred when the Mobile Host moves from one cell to another in case of Cellular IP networks, while it must be transferred from one subnet to another in case of Mobile IP networks [10].

4 The Proposed Model

The procedure of HC context establishment is a long procedure and if it will be repeated each and every time the MN connects to a new Base Station, there will be wastage of the time for both the Mobile Node and the network in general. We propose the following scenario for HC Context transfer.
As we studied in section 2.2.1 there are two types of handoff in Cellular IP network. We assume that the Mobile Node in our proposal performs soft handoff; therefore before the MN connects to the new Base Station (through handoff operation) it will send the soft packet. When the MN sends this packet to the new Base Station, it is approaching to; it will (according to the proposal) append the Context it has established, as we have discussed in section 3, to this soft packet. Now the soft packet consists of request for registration in this cell and HC Context established by the MN. Upon receiving the soft packet, the Base Station will get the HC Context and it may optionally send back an acknowledgment to the MN

4.1 Model as a transition diagram

The Cellular IP model is analyzed for the purpose of HC Context Transfer in Cellular IP networks. The operations are presented as a transition diagram. Four bits are used to represent each state in this transition diagram that consists of fourteen states. Figure 6 shows the transition diagram that describes the proposed model. In this state diagram, BSh refers to the Base Station in which the MN resides at present. BSf refers to the Base Station to which the MN is approaching. In the following table, the meaning of each input signal that appears in the transition diagram is shown

Table 4. Input strings in transition diagram

Input String	Operation that it represents
0000	Information about the network sent to the MN from the BSh
0001	Packets sent from BSh to MN and the signal is strong
0010	Packets sent from BSf to MN and the signal is strong
0011	Registration request from MN to BSh
0100	Registration reply from BSh to MN
0101	Data sent from MN to CN
0110	Data received from MN by BS
0111	Full packets are sent from BS to MN
1000	Partially compressed packets are sent from BSh to MN
1001	Fully compressed packets are sent from BS to MN

4.2 Explanation of the transition diagram

After the MN is switched on and the input string is (0000), the transition diagram will move to the state Q1. From this point, the MN expects two input stings in the transition diagram. If the input string is (0001), which refers that the signal strength coming from the BSh is strong enough the MN will send a registration request to BSh to register itself in this Base Station, the another input signal that the MN expects is (0010) which refers that signal strength coming from a neighboring cell is stronger than the signal strength coming from BSh. If we follow the transition diagram according to the first expected input string, we will see that after sending the registration request to the BSh the later will respond by sending a registration reply, and then the MN will start its session by sending data packets to a CN, in another cell for example. These data packets will be sent through the BSh which is responsible in Cellular IP network for routing these packets to their final destination (as discussed in section 2.2.1). When the data packets reach to the BSh, Header Compression operation will take place through three state Header Compression scheme discussed in section 3.2. This stage will include three state in the state diagram starting in state Q6 and ends in state Q8. During these three states the Header Compression Context will be established as well in both sides compressor (BSh) and de-compressor (MN). After establishment of the HC Context in both sides, this Context should be transferred to another Base Station (BSf) in case of the MN handoff to this Base Station. Therefore, from this point in the transition diagram if the input string is (0010) again the MN will move to state Q10 and the MH sends a rout update packet to BSf and the state diagram reaches state Q11. From this point if the Context has been established (which is the case being explained now), MH will send a registration request to BSf, along with this registration request, then it will append the established Context and the state reached now is Q12. If the Context is not transferred in the way that we proposed, then input signal will be (1110) and the Context establishment procedure between BSf and MN will take place again which will cause waste of the time. After that BSf will send back a registration request to the MN and the registration operation has been completed now. MN is connected to BSf and can continue its session directly by sending data packets. On the other hand if we follow the other side of the transition diagram in which the input signal is (0010), the MH will send a route update packet to BSf reaching the state Q11, while HC Context hasn't been established yet. Therefore, HC procedure will take place through three state Header Compression scheme. After this procedure finishes, state diagram will move from state Q9 to state Q12 and continue as explained previously.

From this transition diagram the benefit we can achieve in Cellular IP system is that we avoid repeating HC Context establishment, if we apply the proposal stating that we append the established Context to the registration request sent to BSf.

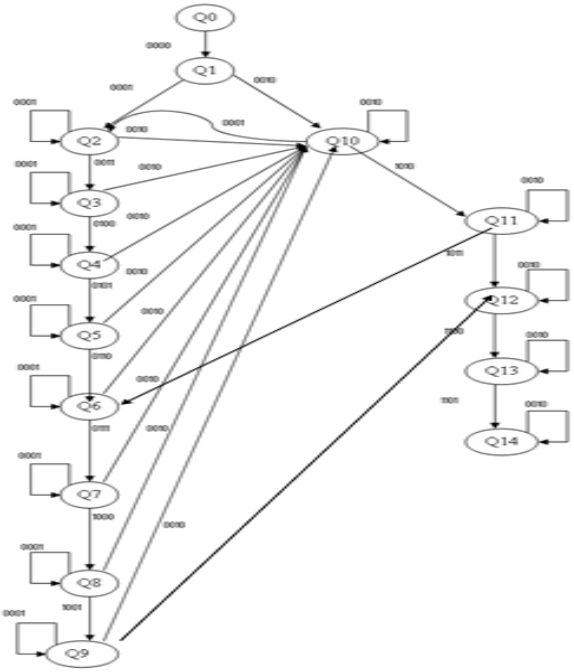


Fig.6.Transition diagram for Cellular IP

5 Experimental Studies

Our experimental studies include the study of probability of error during Header Compression and the variation of packets transfer with Context Transfer and without Context Transfer. The experiment was carried out to observe the time with Context Transfer and without Context Transfer. Finally we carried out the experiment to study the effect of Header Compression on bandwidth using NS-2 simulator.

5.1 Probability of error during header compression

There is always a possibility of the error that may occur during Header Compression operation

through which the establishment of Context takes place.

Probability that no error occurs is $P_n = (1 - \alpha)$ Where α is the Bit Error Rate (BER) (probability of error in one bit).

Probability that there is no error in one byte is $(1 - \alpha)^8$

Probability that there is error in one byte in the packet is $P_b = 1 - (1 - \alpha)^8$

By applying these two probabilities for compressed packet and uncompressed packet the following results are obtained. The result is explained by a bar diagram (fig. 7).

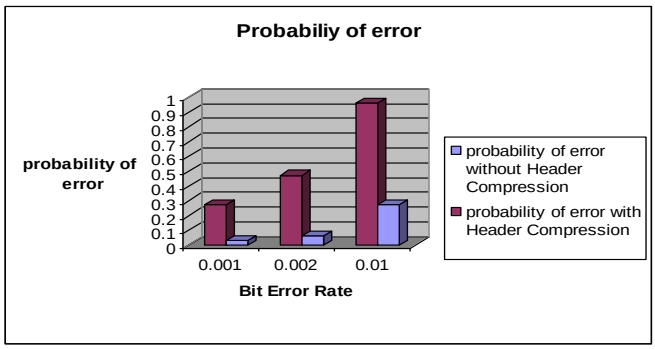


Fig.7. Probability of error for various BER

5.2 Packet transfer time and packet size with Context Transfer and without Context Transfer

It can be noted from the transition diagram that the Context establishment takes three states Q6, Q7, Q8 to complete this operation, while transferring the established context consumes only two states Q10, Q11. Suppose that the Cellular IP model has data transfer rate of 128 kbps, and data traffic is being studied in the direction

BS ⟶ MN. The transfer time is observed for varying packet sizes generated randomly. The experimental results are shown in Table II.

Table 2. Packet size with transfer time

Session No.	Packet No.	Packet size without context transfer (in bytes)	Transfer time for packet (in msec)	Total transfer time without context transfer msec	Packet size with context transfer (in bytes)	Transfer time for packet (in msec)	Total transfer time with context transfer (in msec)
1	P1	800	50	281.1	306.66	19.6	95.4
	P2	450	28.1		11.5	7.96	
	P3	1500	93.7		575	35.93	
	P4	1000	62.5		383.33	23.95	
	P5	750	46.8		287.5	17.96	
2	P1	1300	81.2	33.6	498.33	31.1	13.17
	P2	500	31.2		191.66	11.9	
	P3	700	43.7		268.33	16.7	
	P4	850	53.1		325.83	20.3	
	P5	1350	84.3		517.5	32.3	
	P2	950	59.37		364.16	22.76	
	P3	600	37.5		230	14.37	
	P4	700	43.75		268.33	16.77	
	P5	1550	96.87		594.1	37.13	

For clear observation the results are shown in figures (8.a.1 and 8.b.1)

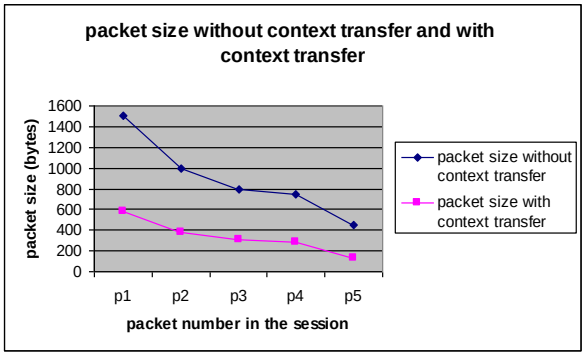


Fig. 8.a.1. session 1

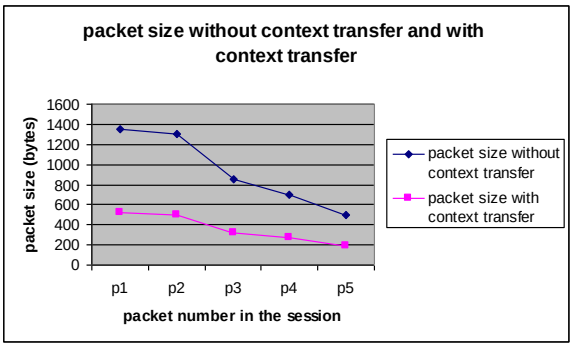


Fig. 8.b.1. session 2

We can note the following, from the results.
1- Each session that takes place in the direction BS ⟶ MN consists of many packets of different sizes and that depends on the data exchanged between MN and BS. In the experiment, the packets are randomly generated.

2- We can observe from table 2 that packet size is bigger when the HC context has not been transferred. It means that the MN is connected to BS and exchanging data with it but without Header Compression i.e. the context has to be established during Header Compression operation. Therefore while the Headers haven't been compressed yet, the packet size is big and transfer time for each packet is also big.
3-When the MN connects to BSf (for example), and the HC context has been transferred with the registration request (explained previously), we see that the MN has to send the compressed packets only to this BS where the context has already been sent and the BS has to decompress the arrived packets only. As a result the packet size is smaller and obviously transfer time is also smaller.
4- The time saved in this case is the time required for sending the established context. For example if we take session 1 the total time required for transferring all its packets without context transfer is 281.1 (msec), while the time required for sending its packets in with context transfer is 95.4 (msec); therefore we saved the time 281.1-95.4= 185.7 (msec) which is the time taken for context establishment and context transfer.
5- Our experiment includes context establishment time and context transfer time, but since we are concerned with packet transfer time we haven't mentioned their values in this paper.

5.3 Effect of header hompression on bandwidth

To study the effect of Header Compression on Bandwidth we used NS-2 simulator, where we can build the suitable model for Cellular IP and a part from the model was taken which consists of three Base Stations and two Mobile Nodes as it is shown in figure 9.
In this model we calculate the consumed bandwidth using the NS-2 simulator in two cases
Case I without Header Compression.
Case II with Header Compression.

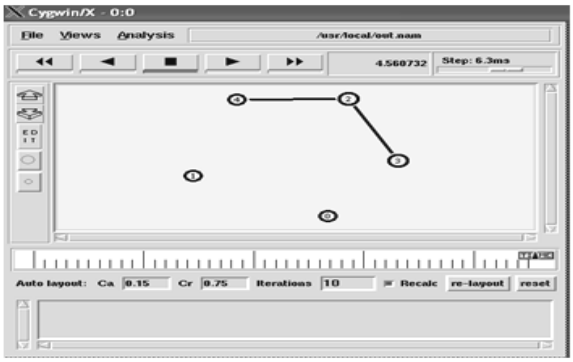


Fig.9 Cellular IP Model

Where
2, 3, 4 Base Stations
0 MN connects to the Base station 3
1 MH connects to the Base Station 4

Case I: Consumed bandwidth in case of not using Header Compression

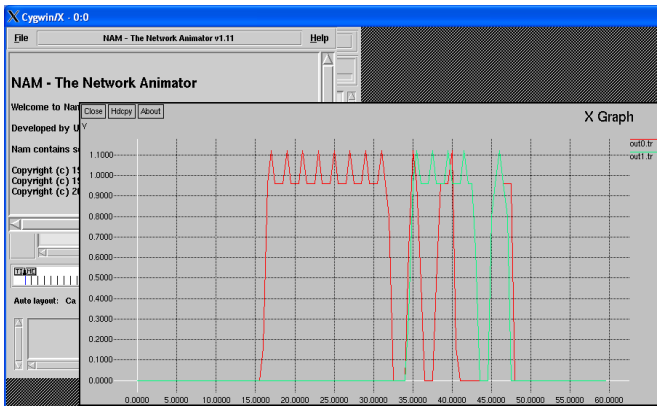


Fig.10.Consumed bandwidth without Header Compression

Case II: Consumed bandwidth with Header Compression

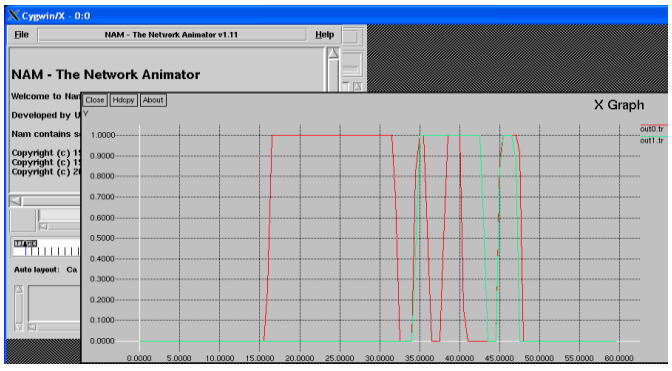


Fig.11.Consumed bandwidth with Header Compression

We found the following results represented by the graphs in fig. 10 and 11.
In Cellular IP network model that we established using NS-2, we have two MN connected to two Base Stations; therefore we have two communication channels and two traffic signals according to the packets sent by every MN through every channel. X-graphs are used to represent packets traffic in these two channels and each X-graph consists of two axis, the horizontal one represents the time taken to send packets through the traffic channel (in seconds), while vertical axis represents bandwidth utilized during data packets transmission (in bytes). From these two X-graphs we can observe the following points:
1- In NS-2 simulator we didn't take care of the fields that are removed from the headers and became as HC context, we just concentrated on the compressed packet with less size than the uncompressed packet with big size; therefore when the MN sends uncompressed packets through traffic channel the maximum bandwidth consumed is 11000 Kbytes/sec (figure10) and there is variation in this amount depending on data packets are sent.
2- We were sure before performing this experiment that the bandwidth consumed when the MN sends compressed packets (expressed as packets with less size in NS-2 simulator). This fact is being proved using the simulator. The thing observable from the second two X-graphs is that the consumed amount of bandwidth becomes stable for a period of time for a number of packets and the maximum amount of bandwidth consumed in this case is 10000 Kbytes/sec (fig.11).

6 Conclusion

Through our study of this model for Header Compression Context Transfer in Cellular IP in which avoidance of HC Context reestablishment was done by suggesting transferring this Context to the new cell towards which a MN is approaching, we observed that the proposed model can improve the QoS in Cellular IP networks. Header Compression is considered to be an important activity and should be provided to the users in Cellular IP networks. The most important thing is the transfer of the Context of this service from one sub-network to another for the user to continue using the services and to save the resources of the network such as the time and the bandwidth.
From experimental results regarding bandwidth it is observed that there is a good improvement for bandwidth in Cellular IP network with Header Compression which will lead us to Context establishment and later on transferring this Context. This improvement can be noted from the comparison between the consumed bandwidth in case of using Header Compression and in case of not using Header Compression with the associated X-graphs. It is also observed from the result that how the time taken in transferring the packets is reduced with the proposed model. Probability of error is also minimized. We hope that the proposed model will be useful for its implementation in Cellular IP networks.

References

[1] Asoke.K.Talukder "Mobile computing technology, applications and services creation". Tata McGraw-Hill Publishing Company Limited, 2005, pp 7-23
[2] http://EFFNET.com accesses on 25th –May-2021
[3] "Simulation of Mobile Internet protocol" www.it.iitb.ac.in/xnet/mobile_ip.
[4] A.T.Campbell, J.Gomez, S.Kim, A.G.Valko, Chieh-Yih Wan, Z.R.Turanyi, "Design, Implementation, and Evaluation of Cellular IP", IEEE Personal Communication, Volume 7, August 2000, pp 42-49.
[5] V. Jacobson/1/, "Compressing TCP/IP Headers for Low-Speed Serial Links", RFC 1144, February 1990.
[6] Ha Duong, Arek Dadej, Steven Gordon, Institute for Telecommunication Research, University of South Australia, "Transferring Header Compression Context in Mobile IP Networks". www.itr.unisa.edu.au/~sgordon/doc/transferring-witsp2003.pdf
[7] Marco Carli, Alessandro Nrli, Alessandro Neri, Andrea Rem, "Mobile IP and Cellular IP Integration for Inter Access Network Handoff", IEEE International Conference on communication, Voliume8, 11-14 June 2001, pp 2467 - 2471
[8]Eriko Nurvitadhi, Ben Lee, Chansu Yu, Myungchul Kim, "Adaptive Semi-Soft Hand off for Cellular IP Networks", www.academic.csuohio.edu/yuc
[9]Jeremy Lilley, Jason Yang, Hari Balakrishnan, Srinivasan Seshan, MIT laboratory for computer science, "A unified Header Compression Framework for Low-Bandwidth Links". www.sigmobile.org/awards/mobicom2000-student.pdf
[10] Changli Jiao, Loren Schwiebert, Golden Richard "Adaptive Header Compression for Wireless Networks". Proceedings of 26th Annual IEEE Conference on Local Computer Networks, 14-16 Nov. 2001, pp 377 – 378.
[11] M. Georgiades, "Context Transfer support for IP-based mobility management", CCSR Awards for Research Excellence 2004, Surrey, UK.
[12] Anton M. and Viatcheslav P. S., On the issue of IP header compression application in high voltage digital power line carrier channels, 2016 International Siberian

Conference on Control and Communications (SIBCON), 2016.

[13] Mate Tomoskozi, Patrick Seeling, Péter Ekler, and Frank H. P. Fitzek, Performance evaluation of network header compression schemes for UDP, RTP and TCP, Periodica PolytechnicaElectrical Engineering and Computer Science,

Demonstration of Denial-of-Service Attack on an Internet-of-Things Device

Samet Tonyalı[1*]

[1]*Abdullah Gül University, Kayseri, TURKEY*

Abstract

Wireless communication area is one of the areas that improved fast with the recent developments in devices with wireless capabilities. This fast pace resulted in the creation of new products that started to take an important role in our lives immediately. A vast majority of these products belong to the group called IoT. One of the hot research topics about IoT devices is communication security. In this paper, we investigate and demonstrate Denial-of-Service (DoS) attacks on a ZigBee device. We summarize different types of DoS attacks that can be applied on ZigBee in the literature, give information about the hardware we used for the attack and present our attack strategy. Our proof-of-concept work showed that some ZigBee modules are vulnerable to DoS attacks, and serious countermeasures should be taken before integrating them into real world applications.

Keywords: *IoT, ZigBee, DoS Attack, Security*

1 Introduction

In today's world, data has an integral part. It is used in almost all aspects of human life in order to improve our lives. Because of this, data collection process and availability of this data are important topics to consider. Internet-of-Things (IoT) refers to the collection of *smart* devices that can collect real world data via the sensors built in them and can integrate this data to the machine (virtual) world with limited resources available for them.

To make this integration possible, IoT devices need a wireless communication mechanism built in them as it is not feasible to use a wired communication for exchange of data, because some IoT devices are put inside human body. IoT includes different types of devices so this creates a need for a mechanism that can work on all of these different platforms, work with the limited resources available and make the data exchange process efficient. ZigBee is a solution that is proposed to fill this gap. It is a framework that works on the Network and Application Layer. ZigBee networks are created for data exchange. These networks consist of three components, which are ZigBee coordinator (ZBC) that is responsible for the control of the network, ZigBee router (ZBR) that routes traffic inside the network and ZigBee end device (ZBE) that collects data and exchanges data with other devices in the network. Only one ZBC can be available in a ZigBee network, whereas more than one ZBR and ZBE can be put in the same network.

As ZigBee is a newly established framework with the invention of IoT devices and the data exchanged in the network needs to be kept private, its vulnerabilities are investigated and tested by performing different attacks. One of the popular attacks is Denial-of-Service (DoS) attacks (Figure 1. A classical DoS attack). These attacks aim to kill or cripple a target so that it cannot provide service to the real users. In this paper, we present a DoS attack that is performed successfully on a ZigBee network.

The rest of the paper is organized as follows. Section 2 Related Work summarizes the similar works to our study. In Section 3 Attack Hardware and Network Information, we present the details of the attack hardware used and targeted network. In Section 4 Our Attack Strategy, we explain our attack strategy and present the result. Finally, we conclude the paper in Section 5 Challenges Encountered.

* Corresponding author: samet.tonyali@agu.edu.tr

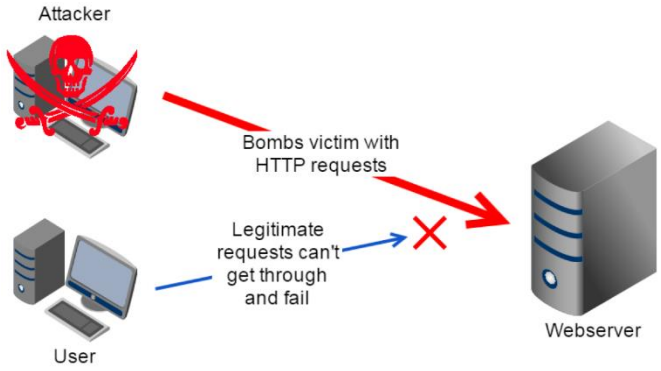


Figure 1. A classical DoS attack

2 Related Work

Vulnerabilities of the IoT devices are exploited in many different ways to cause DoS in an IoT network in literature. These attacks can be examined in five different methods: replay, jamming, flooding, black hole, and energy depletion.

Replay attacks aim to take advantage of the replay protection mechanism of IoT networks. Mechanism is used for DoS when a packet with a really high sequence number is sent to a receiver, causing other packets to be dropped by the receiver as they are considered as *replay*. One of the ways to do replay attack is possible if AES-CTR is used for encryption because of its counter mechanism [72][77][78]. Another way is using KillerBee Framework, a framework that attacks ZigBee devices to crack their key to perform this attack [68]. A ghost attack mechanism that depletes the energy of the device first and then performs replay attack when device counters are reset is also presented in the literature [80].

Jamming attacks try to cause DoS by cancelling transmissions between the sender and the receiver. Three different types of jamming attacks are discussed in literature, which are wide-band jamming, that jams all channels, selective jamming, that jams channels only if there is 802.15.4 traffic available, and message-specific jamming, that jams specific messages rather than all traffic [69][75][80].

Another type of DoS attack that can be done is flooding attacks. These attacks flood a device with a specific message to prohibit the device from answering other devices. A variant of this attack uses acknowledgement (ACK) messages for this purpose [69]. SYN messages can also be used for flooding, too [71] [80]. Another method takes advantage of the CSMA/CA's backoff mechanism and causes other devices to backoff every time they try to access [69][80]. Guaranteed time slots (GTS) can also be targeted for flooding purposes [78]. KillerBee framework is able to do flooding attacks, too. Framework can create random MAC addresses and uses these addresses to associate with a network causing overflow in the network [68][77].

Attackers can act as a black hole to cause DoS, too. If an attacker takes over a ZBR in the network, it can prevent the forwarding of the packets that are coming to this router and eliminate these packets like a black hole [69][71].

The last method that is used in the literature is energy depletion. These attacks try to drain the battery of the devices in the network to eliminate them. One of the attacks can drain the battery of ZBEs by impersonating a ZBC or ZBR and sending broadcast messages [76], whereas another one can focus on a single node in the network and deplete its energy [69][71][80].

3 Attack Hardware and Network Information

In this work, we utilized several different hardware devices. Our target network will be constructed from XBee 802.15.4 Series 1 radios (Figure 2) which implement the IEEE 802.15.4 standard at 2.4 GHz although variations exist at other frequencies and with proprietary modifications to the ZigBee protocol. In order to perform the attack we plan to leverage the KillerBee framework. The KillerBee framework is a suite of security utilities designed for use with 802.15.4 and includes custom firmware for certain supported devices. This firmware helps facilitate a variety of attacks on IEEE 802.15.4-based networks. Specifically, we used the Atmel RZ RAVEN USB stick which is essentially a 2.4 GHz transceiver supported by the KillerBee framework. We needed this device because some of the attacks rely on exploits at the lowest layers of the network stack which cannot be accessed closed devices such as the XBees.

Figure 2. XBee 802.15.4 Series 1 Radio

In addition to these wireless devices, we also utilized an Arduino Uno microcontroller. The Arduino used as the source of data on our target network. It was programmed to count from 0 to infinity and broadcast each iteration on to the network. By monitoring another node on the network, we were then able to easily verify if an attack had been successful by checking to see if there were any gaps in what should be sequential data.

In order to utilize the Atmel RAVEN USB dongles, custom firmware had to be loaded into their flash memory. Unfortunately the RAVEN dongles do not posses In-System Programmer (ISP) chips. The only interface available for loading new firmware on to the devices is through the JTAG debugging header. In order to utilize this interface, a specialized AVR programmer is necessary. The AVR Dragon was selected as an affordable option that had been endorsed by the community for specifically this application. The AVR dragon, along with a few adapters to deal with physical connector differences, can be used to modify the firmware of any AVR device with an exposed JTAG interface such as out 802.15.4 dongles. The KillerBee project provides pre-compiled firmware for the specific devices used in this project, so modifying the firmware is a simple matter of invoking a single command to push a hex file into the programmer and onto the device. You can find a setup to flash the hex file into the device in Figure 3. If the flashing is successful the LED light changes as in Figure 4.

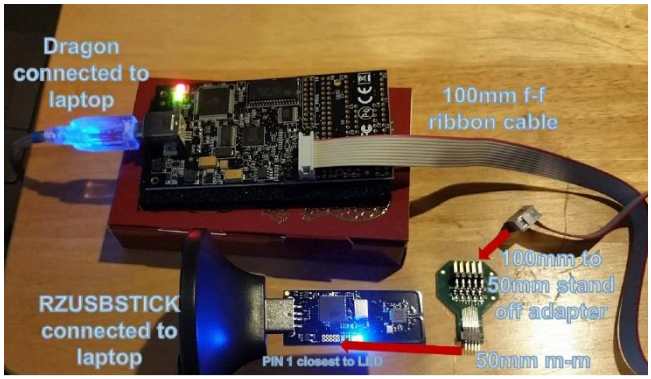


Figure 3. The connections while flashing KillerBee firmware into the USB stick (By courtesy of [73])

Figure 4. LED lights before and after flashing the firmware (By courtesy of [73])

4 Our Attack Strategy

We aim to perform a DoS attack as shown in Figure 1 on a network that consists of two XBees. One of the XBees is the ZBC of the network while the other is a ZBE. The devices run a very simple application. The ZBE sends successive numbers from 0 to infinity to the ZBC. We want the ZBC to be blocked receiving as many packets as possible. To this end, we use the disassociation mechanism in ZigBee network layer implementation. The disassociation can be initiated by either the ZBE or the ZBC. The procedure for the disassociation initiated by the ZBC is given in Figure 5. In beacon-enabled mode, if the ZBC attempts to disassociate the ZBE, it sends a beacon to the ZBE, which states that there is

pending data to read at the ZBC. The ZBE responds with a data request message. The ZBC *ACK*s this request and then sends the disassociation notification to the ZBE. The ZBE *ACK*s this message and its MAC layer notifies the network layer that the device was disassociated from the network.

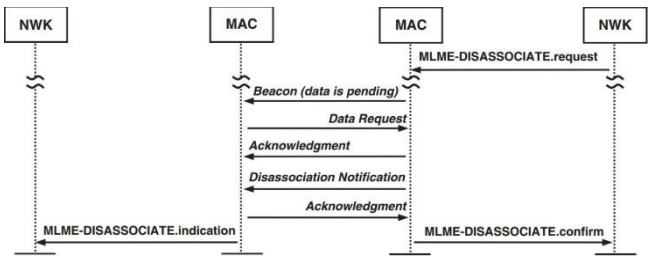


Figure 5. The procedure for the disassociation initiated by the ZBC in beacon-enabled mode (By courtesy of [70])

In our work, we use KillerBee framework on the Atmel RZ RAVEN USB stick (Figure 6) to spoof the ZBC. The stick sends a flood of disassociation messages to the ZBE as if it is the ZBC. We use the *zbdisassociationflood* tool which is an implementation of disassociation procedure in BeekeeperWIDS framework [74] to perform the attack. The devices communicate over plaintexts and do not provide authentication. Since the messages are not authenticated, the ZBC can be spoofed easily. If the messages are authenticated, the *zbdsniff* tool in KillerBee framework can be used to capture the key. Then, the same attack can be performed as explained.

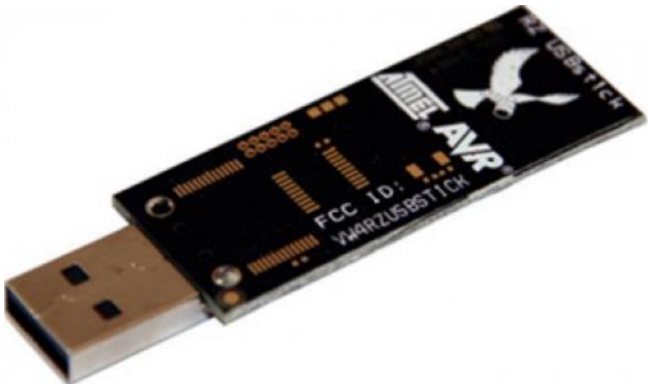


Figure 6. Atmel RZ Raven USB Stick

We gave these parameters to the tool: The communication tool, PAN ID, coordinator device ID, end device short ID, end device ID and number of disassociation request floods. We started the end device to send the consecutive numbers and then, we ran the tool. The ZBC stopped to receive data packets from the ZBE during the attack as given in Figure 7.

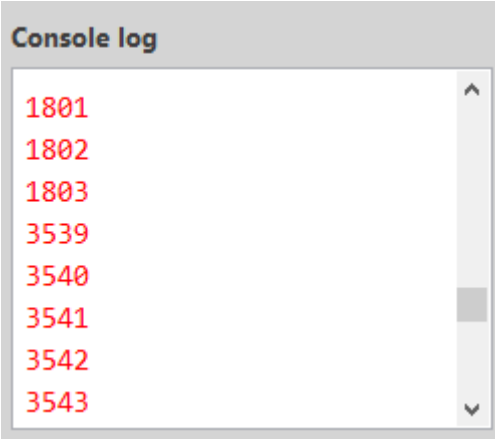


Figure 7. A screenshot that demonstrates that our attack worked

5 Challenges Encountered

Although KillerBee and BeeKeeperWIDS frameworks include useful tools, documentation for both is not sufficient. We sometimes needed to delve into the python code in order to understand what type of data is required for some parameters. In addition. some of the tools did not meet our needs. Therefore, we needed to utilize other tools to perform a DoS attack.

References

[68] Wright, J. "Killerbee: practical zigbee exploitation framework." *In 11th ToorCon conference*, San Diego, vol. 67. 2009.

[69] Radmand, P, Marc D, Jaipal S, Joan A, Alex T, Stig P, and Simon C. "ZigBee/ZigBee PRO security assessment based on compromised cryptographic keys." *In 2010 International Conference on P2P, Parallel, Grid, Cloud and Internet Computing*, pp. 465-470. IEEE, 2010.

[70] Farahani, S. "ZigBee wireless networks and transceivers." *Newnes*, 2011.

[71] Doddapaneni, K, and Arindam, G. "Analysis of Denial-of-Service attacks on Wireless Sensor Networks using simulation." *IT Security for the Next Generation-European Cup 2011* (2011).

[72] Sastry, N, and David, W. "Security considerations for IEEE 802.15. 4 networks." *In Proceedings of the 3rd ACM workshop on Wireless security*, pp. 32-42. 2004.

[73] SecuritySnapse. “Fun with Zigbee Wireless - Part V”. http://securitysynapse.blogspot.com/2015/12/fun-with-zigbee-wireless-part-v.html (15.03.2021)

[74] River Loop Security. “ZB Disassociation Flood Tool”. https://github.com/riverloopsec/beekeeperwids/blob/master/demo/zbdisassociationflood (15.03.2021)

[75] O'Flynn, Colin P. "Message denial and alteration on IEEE 802.15. 4 low-power radio networks." *In 2011 4th IFIP International Conference on New Technologies, Mobility and Security*, pp. 1-5. IEEE, 2011.

[76] Vidgren, N, Keijo H, Jose L P-A, Juan J R-S, and Pekka T. "Security threats in ZigBee-enabled systems: vulnerability evaluation, practical experiments, countermeasures, and lessons learned." *In 2013 46th Hawaii International Conference on System Sciences*, pp. 5132-5138. IEEE, 2013.

[77] Stelte, B, and Gabi D R. "Thwarting attacks on ZigBee-Removal of the KillerBee stinger." *In Proceedings of the 9th International Conference on Network and Service Management (CNSM 2013)*, pp. 219-226. IEEE, 2013.

[78] Amin, Y M, and Amr T. A. "Classification and analysis of IEEE 802.15. 4 MAC layer attacks." *In 2015 11th International Conference on Innovations*

[79] *in Information Technology (IIT)*, pp. 74-79. IEEE, 2015.

[80] Cao, X, Devu M S, Yu C, Z Y, Yang Z, and Jiming C. "Ghost-in-zigbee: Energy depletion attack on zigbee-based wireless networks." *IEEE Internet of Things* Journal 3, no. 5 (2016): 816-829.

A Novel Local Cross T Pattern (LCTP) for Facial Image Recognition

Arif Metehan Yıldız[1*], Türker Tuncer[2]

[1]*Ardahan University, Ardahan, TURKEY*
[2]*Fırat University, Elazığ, TURKEY*

Abstract

Local descriptors have been widely used in facial image recognition and it is one of the most effective method for face identification. In this paper, a novel local cross T pattern (LCTP) is proposed. LCTP uses 7 x 5 size of non-overlapping blocks to extract 512 dimension of feature. In the LCTP, signum function is used to binary feature coding. A novel facial image recognition method is presented using LCTP. The proposed method consists of pre-processing, feature extraction and classification phases. 5 x 5 size of median filtering and singular value decomposition (SVD) are used. In order to feature extraction, LCTP is used. Two classifiers are used in classification phased. These are Linear Discriminant Analysis (LDA) and Support Vector Machine (SVM). To evaluate performance of the proposed LCTP based facial image recognition method, AT&T facial image dataset are used. The experimental results clearly demonstrated that, the proposed LCTP based facial image recognition method has well recognition ability.

Keywords: *Local Cross T Pattern, Face Recognition, Pattern Recognition, Biometrics, Machine Learning*

1 Introduction

Local descriptors have been commonly used for textural and facial image recognition and local binary pattern (LBP) is the first and widely used local descriptor in the literature. LBP was presented in 1996 by Ojala. LBP is well discriminator for texture image. It uses 3 x 3 size of neighborhood matrix and signum function to feature extraction. Histogram of the image is used as feature in the LBP. Following the recommendation of LBP, many LBP like methods have been proposed in the literature. The reasons for the widespread use of these methods are as follows.

- ✓ They coding easily.
- ✓ They are well discriminator for biometrics and textural images.
- ✓ They have shorter execution time.
- ✓ Several mathematical methods can be used for binary feature coding.
- ✓ They can be easily used with other methods.
- ✓ They can have applied onto real world applications easily [1-5].

Previously proposed local descriptor based facial image recognition method in the literature are given as follows. Liu et al. [6] presented an extended version of the LBP. LBP generally extracts uniform patterns. To extract non-uniform patterns, they applied several kernels on the LBP and these features were combined to obtain final feature. In order to feature dimension reduction, weigthed principal component analysis (WPCA) was utilized. They tested this method using 3 face image datasets. Pillai et al. [7] proposed local diagonal extrema number pattern (LDENP) to facial image recognition. This method used 3 x 3 size of neigborhood block to extract feature of face images. Several databases were utilized as test suite and the authors compared this method the others. Roy and Bhattacharjee [8] suggested a local wavelet energy based face recognition method. This method used local wavelet energy, a LBP like descriptor and CNN (Convolutional Neural Network). Yang et al. [9] presented local ternary pattern (LTP) based facial image recognition method and they used several parameters for

* Corresponding author: a.metehanyildiz@gmail.com

testing performance of LTP. Li et al. [10] used LBP and deep learning together to detect face spoofing. Alelaiwi et al. [11] presented SPT (Steerable Pyramid Transform) and LBP based face recognition method for electronic health application. They used face recognition to login e-health system. Uzun-per and Gökmen [12] used local walsh transfrom to extract salient feature of facial images. Pujol and Garcia [13] proposed 9 principle LBP and they used these for facial image recognition. This method was tested on FERET database. Chakborty et al. [14] presented local quadruple pattern (LQPAT) for facial image retrivial and recognition. LQPAT used 4 x 4 size of matrix to extract salient features of face images. The dimension of feature was 512. Patil et al. [15] used multi block LBP and contourlet transform decomposition (CTD) together to facial image recognition. In this method, Features of LBP and CTD were fused then linear discriminant analysis (LDA) was used feature reduction. Finally, these features were classified using KNN. Also, they used modified laplacian pyramid transform to extract salient features.

2 Background

In this section, the organizational chart will be mentioned along with the motivation and contributions of the article.

2.1 Motivations

In this article, a new textual descriptor is proposed. The proposed descriptor is called LCTP. LCTP micro structure uses 7 x 5 overlapping size of matrix. The most important feature of this method is to achieve a more successful descriptor for facial image recognition by performing the feature extraction process. A feature set of 2 x 256 size is obtained using LCTP. Maximum pooling is used to reduce the feature set. Quadratic kernel SVM and LDA are used in the classification phase.

2.2 Organization

The organization of the rest of this article is given as follows. The LCTP is presented in Section 3, the proposed facial image recognition method is presented in the Section 4, experimental results are discussed in the Section 5, conclusions and recommendations are given in Section 6.

3 The local cross T pattern

In this study, a novel local descriptor is presented. The proposed local structure uses 4 T like pattern. Therefore, this method called as LCTP. 7 x 5 size of overlapping blocks, T patterns and signum function are used to feature extraction in the LCTP. The graphical outline of the LCTP is shown in Fig. 1.

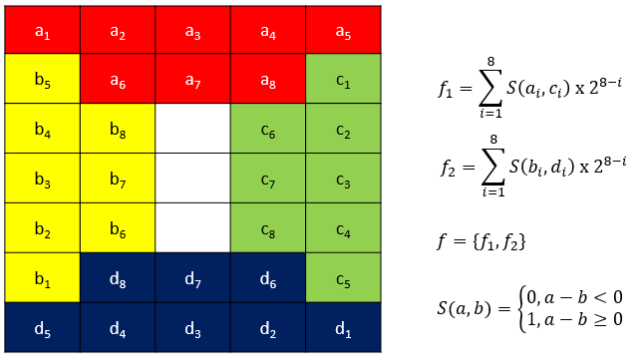


Figure 8. Graphical resresentation of the LCTP.

The mathematical description of the LCTP is given Eq. 1-4.

$$f_1 = \sum_{i=1}^{8} S(a_i, c_i) \times 2^{8-i} \quad (1)$$

$$f_2 = \sum_{i=1}^{8} S(b_i, d_i) \times 2^{8-i} \quad (2)$$

$$f = \{f_1, f_2\} \quad (3)$$

$$S(a,b) = \begin{cases} 0, a-b<0 \\ 1, a-b \geq 0 \end{cases} \quad (4)$$

Where S(.) is signum function and it extract binary features of block, f1 is a and c blocks features, f2 is b and d blocks features and f is final feature.

(a)

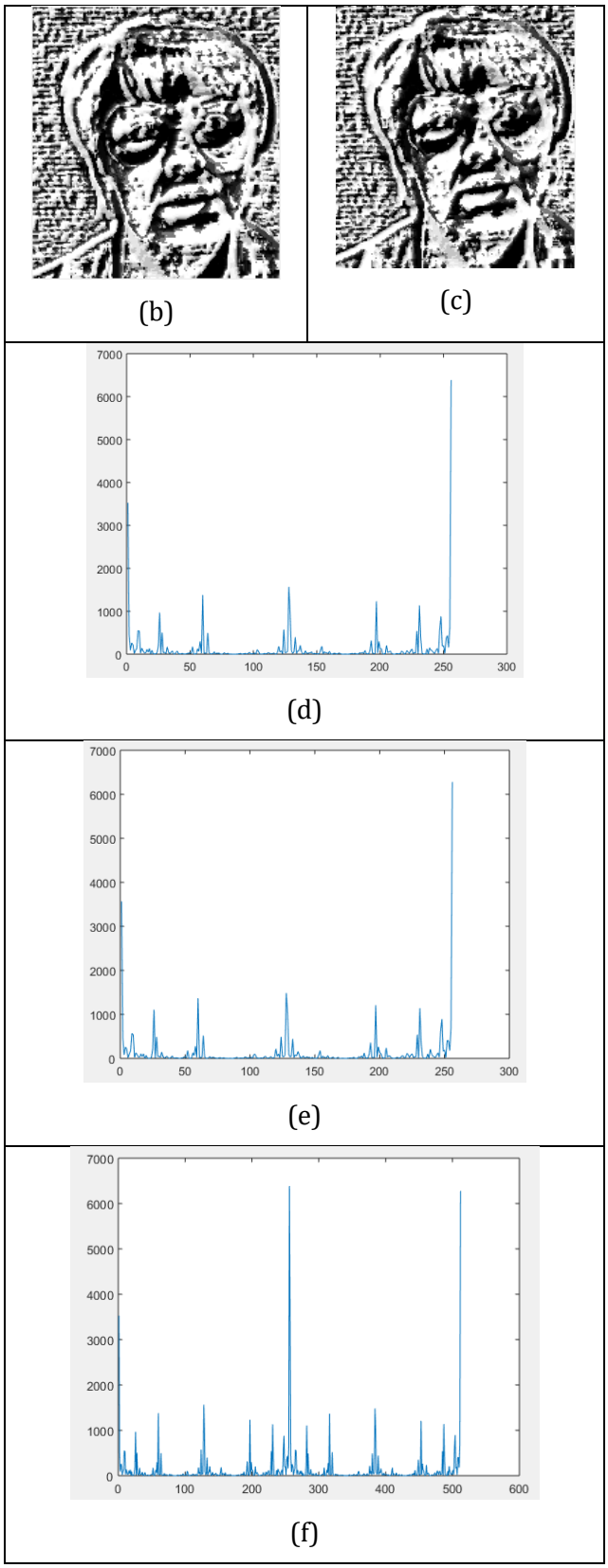


Figure 9. An example about LCTP (a) original face image, (b) first LCTP image, (c) second LCTP image, (d) histogram of the first LCTP image, (e) histogram of the second LCTP image, (f) combined histogram..

4 The Proposed LCTP Based Face Recognition Method

In this study, LCTP based facial image recognition method is presented. This method generally consists of pre-processing, feature extraction and classification phases. In the pre-processing phase, median filtering and SVD are used. The median filtering provides noise reduction. SVD uses 4 x 4 size of non-overlapping blocks and it provides salient and robust feature extraction against geometrical attacks. To create secondary image, SVD is used. LCTP is used to feature extraction. LCTP is used to textural feature extraction. In the classification phase, two classifiers are used and these are LDA and SVM. The block diagram of the proposed method is shown in Fig. 3.

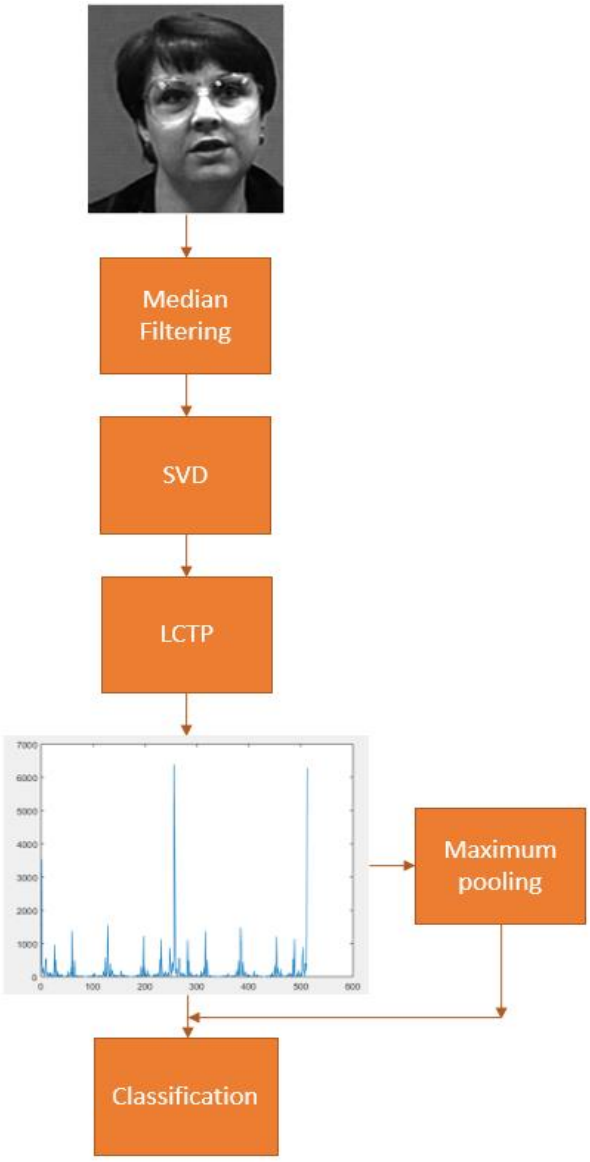


Figure 10. The graphical respresentation of the proposed facial image recognition method.

4.1 Pre-processing

The steps of the pre-processing phase are given below.

Step 1: Load raw face image.

Step 2: Apply 5 x 5 size of median filter onto face image.

Step 3: Divide image into 4 x 4 size of non-overlapping blocks and apply SVD onto each block.

$$SVD(block) = \begin{vmatrix} U_{1,1} & U_{1,2} & U_{1,3} & U_{1,4} \\ U_{2,1} & U_{2,2} & U_{2,3} & U_{2,4} \\ U_{3,1} & U_{3,2} & U_{3,3} & U_{3,4} \\ U_{4,1} & U_{4,2} & U_{4,3} & U_{4,4} \end{vmatrix} \begin{vmatrix} S_1 & 0 & 0 & 0 \\ 0 & S_2 & 0 & 0 \\ 0 & 0 & S_3 & 0 \\ 0 & 0 & 0 & S_4 \end{vmatrix} \begin{vmatrix} V_{1,1} & V_{1,2} & V_{1,3} & V_1 \\ V_{2,1} & V_{2,2} & V_{2,3} & V_2 \\ V_{3,1} & V_{3,2} & V_{3,3} & V_3 \\ V_{4,1} & V_{4,2} & V_{4,3} & V_4 \end{vmatrix} \quad (5)$$

Step 4: Store maximum S values as a pixel values and create secondary image.

4.2 Feature extraction

Algorithm 1. Pseudo code of the calculation pooled feature

Input: Final feature f with dimension of 512
Output: Pooled f (pf) with dimension of 256

```
1: c=1; // c is index counter
2: for i=1 to 512 step by 2 do
3:     vector=f(i:i+1);
4:     pf(c)=max(vector);
5:     c=c+1;
6: endfor i
```

The feature extraction steps are lied below.

Step 5: Apply LCTP to create secondary image.

Step 6: Extract histograms and calculate f1, f2 and f.

Step 7: Use Algorithm 1 and calculate pooled feature.

4.3 Classification

In this section, quadratic kernel and LDA are utilized as classifiers. In the LDA, regularization is diagonal covariance. In the SVM quadratic kernel and one to all multiclass method is used. Also, cross validation folds are 10. f, f1, f2 and pf features are classified in this section.

Step 8: Classify f, f1, f2 and pf features using SVM and LDA.

5 Experimental Results

In this section, the experiments were calculated by using AT&T facial image datasets. AT&T dataset consists of 300 facial images of 30 people. There are 10 images of each individual. The size of AT&T dataset images is 112 x 92 and these images are gray-level. Some of these images are shown as in Fig. 4.

Figure 11. Sample images from AT&T dataset [4].

In order to obtain numerical results, accuracy is used and its mathematical definition lied in Eq. 6.

$$Acc = \frac{True\ predictive\ samples}{All\ samples} \quad (6)$$

Where *Acc* represents accuracy.

The calculated results according to features and classifiers are listed in Table 1.

Table 5. Accuracy values of the proposed method.

Features	LDA	SVM
f1	0.910	0.860
f2	0.873	0.800
pf	0.920	0.933
f	0.960	0.943

Table 1 clearly illustrated that combined and pooled features are more salient and distinctive than f1 and f2.

6 Conclusions and Recommendations

In this paper a novel local descriptor which is LCTP and LCTP based facial image recognition method. LCTP uses 7 x 5 size of blocks and each block is divided in into 4 T like pieces. A novel facial image recognition method based on LCTP is proposed in this study. This method generally consists of pre-processing, feature extraction and classification phases. Median filtering and SVD are used to create secondary image in the pre-processing. In the feature extraction phase 2 x 256 size of feature are extracted. Also, feature fusion and maximum pooling feature reduction methods were used to extract 4 feature vector. These feature vectors are called as f1, f2, pf and f. The dimension of these vectors 256, 256, 256 and 512 respectively. In the classification phase, LDA and SVM were used. The highest accuracy rate was obtained combined feature with LDA classifier and experiments showed that, maximum pooling based feature reduction was increased recognition ability of the f1 and f2 feature sets. Briefly, Table 1 showed that the proposed LCTP based facial image recognition method has high face recognition capability.

In the future studies, novel local descriptors or micro patterns may be proposed and different mathematical kernels such as ternary, quaternary, fuzzy sets based functions can be used in LCTP. Also, SVD like transformations can be used with LCTP for facial image recognition.

References

[1] T. Ojala, M. Pietikinen, D. Harwood, A comparative study of texture measures with classification based on feature distributions, Pattern Recognit. 29 (1) (1996).

[2] Y. Huang, Y. Wang, T. Tan, Combining statistics of geometrical and correlative features for 3d face recognition, in: Proceedings of the 17th British Machine Vision Conference, 2006, pp. 879–888.

[3] L. Qiao, S. Chen, X. Tan, Sparsity preserving projections with applications to face recognition, Pattern Recognition 43 (1) (2010) 331–341.

[4] F. Samaria, A. Harter, Parameterisation of a stochastic model for human face identification, 2nd IEEE Workshop on Applications of Computer Vision December 1994, Sarasota(Florida), http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html (accessed July 1, 2018).

[5] G. Li, J. Kim, Palmprint recognition with Local Micro-structure Tetra Pattern, Pattern Recognition 61 (2017) 29–4.

[6] L. Liu, P. Fiequth, G. Zhao, M. Pietkainen, D. Hu, Extended local binary patterns for face recognition, Information Sciences 358–359 (2016) 56–72.

[7] A. Pillai, R. Sonudrapandiyan, S Sataphy, S. C. Satapathy, K.H. Jung, R. Krishnan, Local diagonal extrema number pattern: A new feature descriptor for face recognition, Future Generation Computer Systems 81 (2018) 297–306.

[8] H. Roy, D. Bhattacharjee, A novel local wavelet energy mesh pattern (LWEMeP) for heterogeneous face recognition, Image and Vision Computing 72 (2018) 1–13.

[9] W. Yang, Z. Wang, B. Zhang, Face recognition using adaptive local ternary patterns method, Neurocomputing 213 (2016) 183–190.

[10] L. Li, X. Feng, Z. Xia, X. Jiang, A. Hadid, Face spoofing detection with local binary pattern network, Journal of Visual Communication and Image Representation 54 (2018) 182–192.

[11] A. Alelaiwi, W. Abdul, M. S. Dewan, M. Migdadi, G. Muhammad, Steerable pyramid transform and local binary pattern based robust face recognition for e-health secured login, Computers and Electrical Engineering 53 (2016) 435–443.

[12] M. Uzun-per, M. Gökmen, Face recognition with Patch-based Local Walsh Transform, Signal Processing: Image Communication 61 (2018) 85–96.

[13] F.A Pujol, J.C. Garcia, Computing the Principal Local Binary Patterns for face recognition using data mining tools, Expert Systems with Applications 39 (2012) 7165–7172.

[14] S. Chakraborty, S.K. Singh, P. Chakraborty, Local quadruple pattern: A novel descriptor for facial image recognition and retrieval, Computers and Electrical Engineering 62 (2017) 92–104.

[15] H. Y. Patil, A. G. Kothari, K. M. Bhurchandi, Expression invariant face recognition using local binary patterns and contourlet transform, Optik 127 (2016) 2670–2678.

A Steganography Algorithm for Digital Image Hiding into Digital Audio

Ali Erdem Altınbaş[1*], Yıldıray Yalman[1]

[1]*Piri Reis University, İstanbul, TURKEY*

Abstract

In this study, a new steganography algorithm has been developed to hide digital images in digital audio files. The developed algorithm basically consists of 3 stages. In the first step, the size of the image to be hidden is calculated and added to the data to be hidden. Size information is used to extract the message. In the second step, the encryption of the message (stego-data) is performed according to Kerckhoffs's principle. During encryption, the image is complemented first, and then the message is encrypted with the One-Time-Pad (OTP) algorithm. In the third stage, hiding is carried out. The presented steganography algorithm is implemented by manipulating a 16-bit digital audio file (cover data) in decimal. Since the change in the sound file is relatively small, the deterioration in sound quality is minimized. The audio file obtained after data hiding is now covered-data. In the stego-data extraction phase, primary, the size information of the image is accessed from the covered-data. Then, the encrypted message is extracted. Then, the encrypted data is decrypted, and the hidden image (stego-image) is reconstructed. The sampling frequency of the audio file used in the developed algorithm is 44100 Hz and the bit depth is 16. In this way, the image that is wanted to be hidden by the small manipulations of the value in each sample is hidden in the sound file in a way that the human ear cannot notice. The experimental results and quality assessments show that the steganography algorithm performed on a decimal base allows the size of the hidden message (stego-data) to increase with small changes in the sound. In addition, it is not possible to realize the effect of this change by examining the spectrogram or the waveform. The MATLAB codes of the developed application and the covered-data are available in the link: https://bit.ly/3dBLzhF

Keywords: *Information Security, Audio Steganography, Data Hiding*

1 Introduction

Steganography is one of the most successful ways of secure communication today. It is carried out by embedding the information to be sent in media that will not cause suspicion by third-party listeners [1]. The first notable example in which the carrier media is sound is the patent obtained by the company Muzak, which produces background music in 1954, about embedding registration information in music records [2]. One of the most interesting examples of the point reached over the years is the use of blockchain technology in important information with hash data to ensure the integrity of the data and to avoid availability problems in a possible connection error [3].

The presented study hides an image into an audio file. The carrier audio file is called cover-data, the image to be hidden is stego-image, and the audio file obtained after the concealment is performed is called as stego-data. Since image hiding is done numerically in decimal, more information is hidden in a relatively small area. In each sample, the units digit is changed instead of the least significant bit.

The study consists of five main sections. In the second section, commonly known steganography studies are examined and the importance of the developed algorithm is explained. In the third section, technical structures of cover-data and stego-data, which are important parts of the study, such as sampling rate, bit depth, chunk, and format are analyzed. The developed steganography algorithm is explained in the fourth section. In the same section, application results such as capacity, quality assessment, and noticeability are presented. The findings obtained are evaluated in the last section.

* Corresponding author: alierdem.altinbas@pru.edu.tr

2 Related Works

Audio steganography refers to the digital manipulation of the cover audio file in the desired way. With a certain algorithm, the message to be hidden can be successfully hidden by changing the digital audio file. Besides, an extraction algorithm should be used to remove the hidden message. Some of the techniques that are widely used and frequently mentioned in the literature are as follows:

Least Significant Bit (LSB): LSB is one of the oldest methods used in audio steganography. Because the application is relatively simple, and the amount of distortion is very low. In this method, in the binary value of any sample of the cover audio file, the LSB value that has the least effect on that sample is replaced with the bit value of the message to be hidden. This method is also preferred as it makes the probability of any bit changing to be 50%. Moreover, the higher the sampling frequency, the more information can be placed in the sound file of the same duration [4].

Echo Hiding: Unlike the LSB, the Echo hiding method has been developed to prevent the change in the cover audio file from being random noise. Sound recordings are recorded as dry as possible. After recording, an environment is simulated using various digital effects such as reverb, echo, or delay in the post-production. The use of these effects causes a numerical change in the sound file. The information to be hidden by the echo hiding method is placed in the cover audio in a way to add an echo effect. This method reduces the possibility of noticing the change with the human auditory system (HAS) in steganalysis [5].

Phase Coding: The information to be hidden in the phase coding method is encoded as phase shifting of the sound. In this method, the fact that phase shifts are less perceptible than noise is essential. Thus, when the cover data and stego data are compared, the signal-to-perceived noise (SPNR) is so high that the difference cannot be heard [6].

Apart from these commonly used methods, there are various approaches in the temporal domain, frequency domain, and wavelet domain [4]. The main purpose of these approaches in steganography is to develop inaudible concealment techniques and to hide more information in fewer areas [7]. The presented study is also related to the algorithm used to increase the capacity. The way to make the deterioration inaudible is related to the technical features of cover data. In this way, both goals (capacity and unnoticeability) are achieved.

3 Digital Audio and Data Hiding

Creating a digital audio file consists of two basic steps; it is first sampled at intervals for a fixed time interval. Afterward, quantization is performed to express the amplitude of each sample [8]. Then the resulting digital audio is basically in .wav file format. The starting point of this file format is Microsoft's file structure with .riff extension [2]. This file structure contains information about the attributes of the raw data, before the data itself, called the chunk. As seen in Figure 1, the data itself starts from the 45th byte.

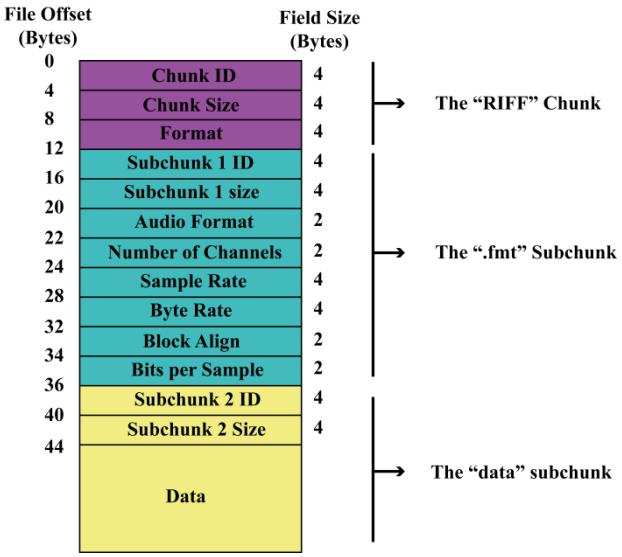


Figure 2. RIFF Chunks

In the data hiding stage, before the data to be hidden in the cover data, an artificial chunk must be added about the size information. In this way, at the stage of extracting the data, it will be possible to learn up to which sample the cover data contains hidden information. Sufficient samples should be reserved to hide the size information to be added. Because in the hiding method, the maximum amount of information that can be hidden in each sample is known in advance. However, of two audio files with the same sampling frequency, the one with the longer duration contains more hiding areas than the other. From this point on, the sampling rate information from the ".fmt" subchunks should be learned before the hiding algorithm starts.

Another piece of information that should be known before the hiding algorithm starts is the size of the message. The hidden message in the presented study is the RGB image. However, all data, including the file

format, not the pixel values of the image are hidden. In addition, expressing an image in binary format as raw data usually takes less space than expressing it in pixels. In this way, the size of the data to be hidden has been reduced. Moreover, if hidden information is captured, even a single bit different causes the binary data not to be reconstructed in the image format.

4 The Proposed Steganography Method

The proposed algorithm consists of two stages: data hiding and extracting. The first stage, data hiding, is shown in the block diagram in Figure 2.

As seen in Figure 2, the size of both the message and the cover data are calculated first. This calculation is necessary not only to hide the size of the message but also to find out whether the size of the message is suitable for the capacity of the cover data. If the capacity of the cover data is suitable for the message, the size information of the message is added to the cover data as a chunk. The message size information is converted to binary value and placed at the last bits of the first 32 samples of the cover data.

After the size of the message is hidden in the cover data, the hiding of the message starts from the 33rd example. The first thing to do at this stage is to get each element of the message complement. Then all elements are expressed as 3 digits. For this, 0 is added to the left of single or two-digit numbers. This process is illustrated in Figure 3.

As seen in Figure 3, the second element of the message is 166. If this element, which is 89 after the 255's complement, is hidden as two-digit '89', it may not be possible to retrieve the data correctly. Because the number of digits of the second element of the message is not known at the data extraction stage. For this reason, it should be known in advance that the data has 3 digits in the data extraction phase.

Original Message	255's Complement	3 Digit Message
56	199	199
166	89	**0**89
247	8	**00**8
45	210	210
215	40	**0**40
110	145	145
239	16	**0**16
188	67	**0**67
49	206	206
63	192	192

Figure 3: Creating the 3-digit message

At the last stage of preparing the message for hiding, encryption is carried out. This is because the main purpose of all cyber security algorithms is confidentiality, integrity, and availability. Although integrity and availability are related to the steganography algorithm in the presented study, encrypting the message to ensure confidentiality increases the level of security [9]. The algorithm chosen for encryption is the one-time pad (OTP) algorithm. Also known as Vernam encryption, the length of the message and the key are equal in this algorithm. The corresponding bits of the message and the key are the inputs of the exclusive-or (XOR) operation. The value obtained at the exit is the encrypted value. It is frequently used because of its simple application and robustness. It is robust even against the attacks of computers with high computing power [10]. However, for a cover audio file that is 1 second long and has a sampling frequency of 44100 Hz, a key of about 44000 bits long is required. To solve the problem of the size of the key, each digit of the message with 255's complement is expressed in 4 bits and LSB and MSB are put into the XOR operation. As shown in Figure 4, the value obtained at the output is re-written to the LSB.

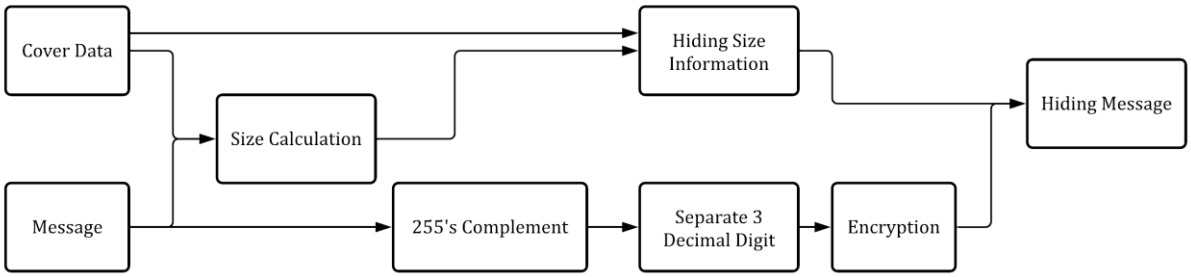


Figure 2: The block diagram of data hiding

255's Complement	Binary Message	Encrypted Message
1	0 0 0 1	0 0 0 1
9	1 0 0 1	1 0 0 **0**
9	1 0 0 1	1 0 0 **0**
0	0 0 0 0	0 0 0 0
8	1 0 0 0	1 0 0 **1**
9	1 0 0 1	1 0 0 **0**
0	0 0 0 0	0 0 0 0
0	0 0 0 0	0 0 0 0
8	1 0 0 0	1 0 0 **1**

Figure 4: OTP Encryption without External Key

In this way, encryption is carried out without the need for an external key. This change, made so that the 4-bit binary number does not exceed 9 in decimal, only decreases and increases the number by one. However, changing only 8 and 9 is enough to ensure the confidentiality of the message if the message is captured by steganalysis methods. Because to ensure file integrity, the encrypted message must be decrypted.

After the encryption process is completed, an *M×1* size matrix elements between 0-9 is obtained. Afterward, the hiding of the message is done in decimals again. Cover data has a depth of 16 bits per sample. This means that the decimal dynamic range of each sample is 65536. Hiding is done by changing the ones digit of the samples. This process is done over the absolute values of the numbers as shown in Figure 5.

Cover Data	Message	Stego Data
-581	3	-58**3**
-460	8	-46**8**
-268	1	-26**1**
-53	9	-5**9**
1	1	**1**
-89	1	-8**1**
-112	9	-11**9**
1	5	**5**
205	1	20**1**
473	9	47**9**

Figure 5: Changing units digit for data hiding

The stego data obtained as a result of hiding the message to the cover data is reconstructed by preserving its original format. The second stage of the proposed algorithm, data extraction, is an operation that takes less time than hiding the data, as shown in the block diagram in Figure 6.

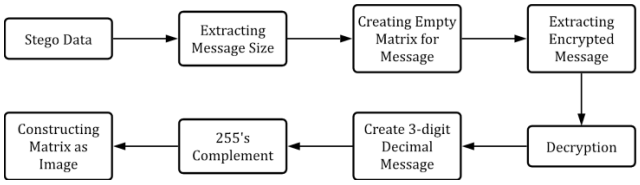


Figure 6: The Block Diagram of Data Extracting

The first step in extracting data is to learn the size of the message. From this point on, it is known up to which sample of the audio file the extraction algorithm will work. To learn the size information, the first 32 samples of stego data are checked. In empty area of 32 bits writes 1 if these samples are odd and 0 if they are even. The 32-bit binary size matrix created shows the message starting from the 33rd bit and continuing up to the number of samples. Next, an empty matrix is created into which the message will be written. The size of the matrix has been learned in the first step. Then, the ones digit of the absolute value of each sample in the audio file containing information about the hidden message is obtained. As a result, a series of numbers containing the numbers 0-9 was formed. These numbers are expressed as a 4-digit binary as when hiding the message. In the next step, the XOR process is performed with MSB and LSB to get the plain message. All transactions up to this stage are exemplified in Figure 7.

Stego Data	Encrypted Message	Encrypted Binary	Decrypted Binary	Decrypted Message
421	1	0 0 0 1	0 0 0 1	1
238	8	1 0 0 0	1 0 0 **1**	**9**
70	0	0 0 0 0	0 0 0 0	0
-160	0	0 0 0 0	0 0 0 0	0
-616	6	0 1 1 0	0 1 1 0	6
-3199	9	1 0 0 1	1 0 0 **0**	**8**
-3311	1	0 0 0 1	0 0 0 1	1
-3064	4	0 1 0 0	0 1 0 0	4
-2738	8	1 0 0 0	1 0 0 **1**	**9**

Figure 7: Extracting and Decryption of the Hidden Message

After the decrypted message is obtained, a 3-digit decimal value is generated. In the example in Figure 6, the encrypted message to be created is 190, 068 and 149. Since the encrypted message is the complement of 255, the plain message is 65, 187, and 106. In this way, information is retrieved from all samples containing a message, and the hidden image is reconstructed.

4.1 Experimental results

Experimental studies show that making the hiding algorithm in decimals increases the data capacity that can be hidden. While performing the payload analysis, scenarios, where 10 to 100 percent of the cover data are used, were examined. According to this, it is thought that the modification made in the cover data will be maximum for each payload rate. In this way, the greatest possible changes were made in each sample. As seen in Figure 8, PSNR remains above 79 dB even if the entire usable area is used.

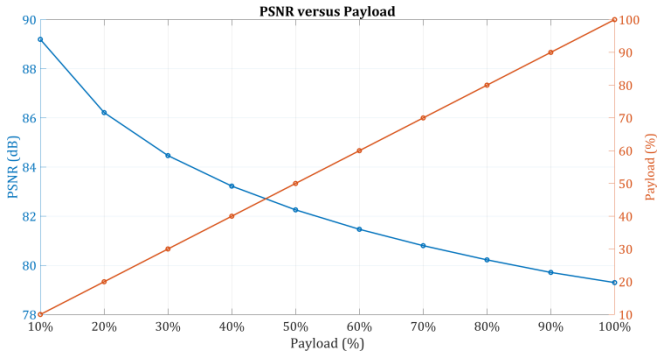


Figure 8: Payload versus PSNR

The size and PSNR values of the cover data used and the prepared message are given in Table 1. The maximum number of bits that can be hidden in the cover file is 21829500 bits. This is because each instance of the cover file is expressed in 16 bits, while the message to be hidden is expressed in 4 bits.

Another experimental result concerns waveforms. The change in normalized amplitudes will be at most 1.3733×10^{-4} %. Therefore, visual inspection of the waveform would be an unsuccessful attempt at steganalysis. Stego audio waveforms with the maximum modification with cover audio are given in Figure 9.

Table 1: Cover Data and Message Size and PSNR Values

Cover Length (Bits)	Message Length (Bits)	PSNR (dB)
882×10^5	2205000	89.194
882×10^5	4410000	86.216
882×10^5	6615000	84.466
882×10^5	8820000	83.224
882×10^5	11025000	82.257
882×10^5	13230000	81.467
882×10^5	15435000	80.801
882×10^5	17640000	80.223
882×10^5	19845000	79.712
882×10^5	21829500	79.298

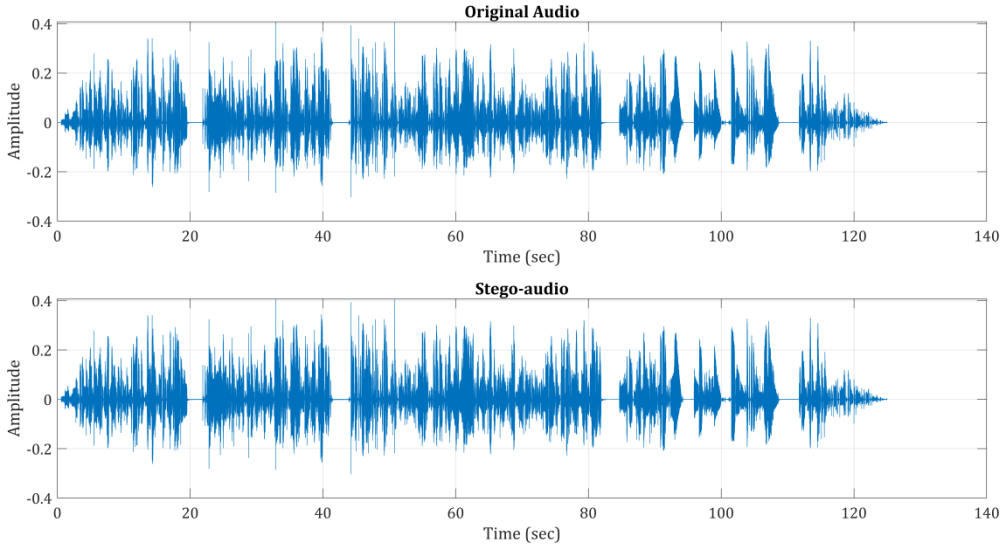
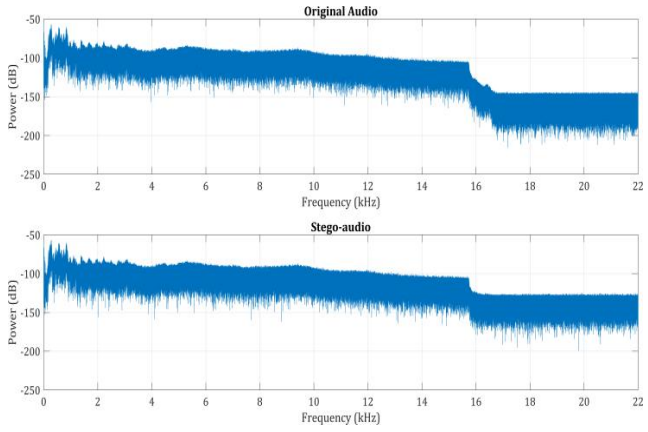


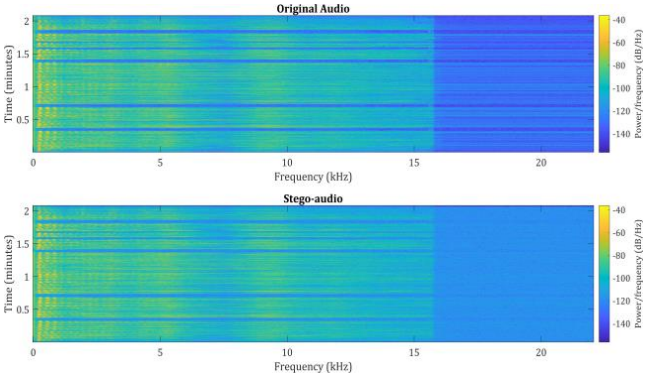
Figure 9: Waveforms of Original Audio and Stego Audio

When the periodograms are examined, it is possible to notice the change in power after 16 kHz in the general structure, as seen in Figure 10 (a), while the main lines are preserved. However, this is not a noticeable change as it is above the hearing threshold of most people [11].

Finally, when the spectrograms are examined, as seen in Figure 10 (b), there is no new pattern due to the hiding process. The effect above 16 kHz seen in the periodogram is more intense in the spectrogram. However, the interpretation that can be made from the spectrogram is that there may be noise in the high frequency components. This irregular and non-patterning structure means that the proposed algorithm cannot be noticed by spectrogram analysis.



(a)



(b)

Figure 10: Periodograms (a) and Spectrograms (b) of Original Audio and Stego Audio

5 Conclusion

In the presented study, a steganography algorithm has been developed in which an image is hidden in an audio file. In the link given in the abstract section, the widely used image of Lena and the song Tom's Diner by Suzanne Vega are used. One of the novelties in the study is that if confidential data cannot be decrypted with the OTP algorithm, the data cannot be generated in image form. In this way, the robustness of the algorithm is increased. Since the work is carried out considering the Kerchkhoffs principle, even if the data is obtained, it is not possible to reach the hidden image unless the key is known.

Considering the experimental results, the presented study increases the data hiding capacity while at the same time keeping the sound quality above a certain level. In this way, it cannot be noticed that a steganography application is made in the cover audio by both auditory and visual methods. In addition, our study guarantees that the PSNR value will be above 79 dB, even if the maximum hiding area is used for each sample. This means that a steganography algorithm with increased hiding capacity has been developed that provides sound quality above a certain level.

References

[1] I. Avcibaş, "Audio steganalysis with content-independent distortion measures," *IEEE Signal Process. Lett.*, vol. 13, no. 2, pp. 92–95, 2006.

[2] P. Dergisi, "Sayısal Ses İçerisinde Gizli Veri Transferinin Kablosuz Ortamda Gerçekleştirilmesi," *Politek. Derg.*, vol. 11, no. 4, pp. 319–327, 2008.

[3] A. H. Mohsin *et al.*, "PSO–Blockchain-based image steganography: towards a new method to secure updating and sharing COVID-19 data in decentralised hospitals intelligence architecture," *Multimed. Tools Appl.*, 2021.

[4] D. Tan, Y. Lu, X. Yan, and X. Wang, "A simple review of audio steganography," in *Proceedings of 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference, ITNEC 2019*, 2019.

[5] D. Gruhl, A. Lu, and W. Bender, "Echo hiding," in *Information Hiding*, 1996.

[6] P. Singh, "A Comparative Study of Audio Steganography Techniques," *Int. Res. J. Eng. Technol.*, vol. 3, no. 4, pp. 580–585, 2016.

[7] N. Cvejic and T. Seppanen, "Increasing the capacity of LSB-based audio steganography," *Proc. 2002 IEEE Work. Multimed. Signal Process. MMSP 2002*.

[8] H. Fastl and E. Zwicker, *Psychoacoustics: Facts and Models*. Berlin, Heidelberg: Springer-Verlag, 2006.

[9] N. Karnataka, "Cryptography for RC4 and 3DES Encryption," no. Icisc, pp. 96–100, 2020.

[10] F. L. Chen, W. F. Liu, S. G. Chen, and Z. H. Wang, "Public-key quantum digital signature scheme

with one-time pad private-key," *Quantum Inf. Process.*, vol. 17, no. 1, pp. 1–14, 2018.

[11] A. P. Benguerel, "Signals and Systems for Speech and Hearing," *Lang. Speech*, vol. 34, no. 4, pp. 381–382, 1991.

Malware Analysis on Android Devices - Dynamic Analysis

Dilek Gökçeoğlu[1], Şengül Doğan[1*]

[1]*Fırat University, Elazığ, TURKEY*

Abstract

As the technology advances, more and more people are using their smart phones and tablets for such daily works as online shopping, internet banking, communication; which is the main reason why vicious people send malware to such smart devices. In this study, which mainly focuses on Android operating system, three types of analysis are touched upon. Static analysis is based on features specified before running the code, dynamic analysis, on the other hand, is based on practice-based features. Network analysis is based on network-based features. While the static analysis of an android application may be based on features gained from the file of "Androidmanifest.xml" or from java byte code, the Dynamic analysis of such applications is based on the analysis of the behaviors of application while running, and on the analysis of the changes made in the code. Network analysis is based on the requests and changes that the application made in the network. The main purpose of this study is to explain how the analysis of malware in Android Operating System is carried out. The structure, the architecture and the file system of Android Operating system are touched upon. In the analysis part, it is explained how to carry out the analyses. The information that is gained after analysis is shared. The study of "sample cade" has been carried out. Finally, the precautions that must be taken are touched upon.

Keywords: *Malware Analysis, Dynamic Analysis, Network Analysis*

1 Introduction

Nowadays the use of smartphone and tablets has become common. These devices are convenient to people to follow the technology. It enables us to easily access commonly used services such as mobile banking, e-book / newspaper reading, games, and social media usage. It is also equipped with various features in many areas such as GPS, Wi-Fi, video calls and more. Ensuring the security of mobile devices that can provide all these features is among the important issues of today. Malware (short for malicious and software) is a pseudonym for disrupting or blocking system operation, collecting information that leads to a breach of privacy or exploitation of the system, gaining unauthorized access to system resources, and other maliciously created software. The three most common types of malware are Viruses, Worms and Trojans [1-3]. These terms are often used interchangeably, but they are actually very different from each other. Each is a type of malware, but they differ in how they infect and affect the systems they will damage. When these malicious software infect mobile devices such as smart phones that we commonly use, it causes the following effects [4].

- Access to message history,
- Accessing location information,
- Stealing password information,
- Stealing bank account information,
- Sending messages to other people as if they were sent by you without your knowledge or permission,
- Making phone calls without your knowledge,
- Using and ending your data package without your knowledge,
- To reduce the usage time of your charge,
- Recording your phone calls and selling them to other people.

2 Android

Android is an open source mobile operating system. It is currently developed by Google and is based on a Linux kernel. The applications are written in Java and converted into a slightly different format known as Dalvik. It is powered by Android Dalvik VM [19, 20].

* Corresponding author: sdogan@firat.edu.tr

2.1 Android application package (APK)

APK literally is known for having the initials of the Android Application Package next to each other. It is a file format used on devices with Android operating system; An APK file is an archive file that typically contains the following files and directories:

• program code,

• certificate

• Manifest files,

• resources.

Apk file is also Jar (Java Archive) and ZIP file. The files and directories contained in an apk file are described below.

2.2 APP components

Application components can be viewed as the basic building blocks of an Android application. Each component is an entry point where the system or user can enter the application. These are;

- Activities
- Services
- Broadcast Receivers
- Content Providers
- Intent

2.2.1 Activities

Activity is the part that creates the user interface that provides interaction between the application and the user.

2.2.2 Services

It is the component that allows applications to run in the background. It does not provide a user interface. For example, it is a service component that allows a user to play music in the background while using a different application.

2.2.3 Broadcast receivers

Broadcast receivers are the component of an application that responds to broadcast messages from other applications or the system itself.

2.2.4 Content providers

A content provider manages the application data stored and shared in the file system, a SQLite database, on the web, or other storage areas that your application can access.

2.2.5 Intent

Three of the four component types - activity, services, and broadcast receivers - are activated by an asynchronous message called an intent. Intents connect individual components to each other at runtime (when the application is running).

3 Dynamic Analysis

Dynamic analysis is performed by running the application on a real device or a simulator [24]. Collected data such as sensitive data access, traffic analysis and insecure requests are analyzed. Dynamic analysis is the stage of examining file read and write operations, open network connections, incoming / outgoing network traffic during the run-time of the application [25].

Dynamic Analysis Platforms:

- AASandbox (Android Application Sandbox)
- TaintDroid
- DroidBox
- Bouncer
- Andrubis
- Droidscope
- AppsPlayground
- Mobile-Sandbox
- CopperDroid
- MOBSF (Mobile Security Framework)

In this research, MOBSF was used as a platform. The malicious application to be analyzed is the apk created by Msfvenom, whose static analysis was performed in the previous study. In the application part of the research, the network traffic of the relevant malicious file and the logs at the time of execution will be examined.

3. 1 Analysis with MOBSF

After the framework was installed, http://localhost: 8000/ was typed in the url part of the browser and the apk file was uploaded to the mobsf page. After the installation, the "Start Dynamic Analysis" tab shown in Figure 5 was selected and the analysis started.

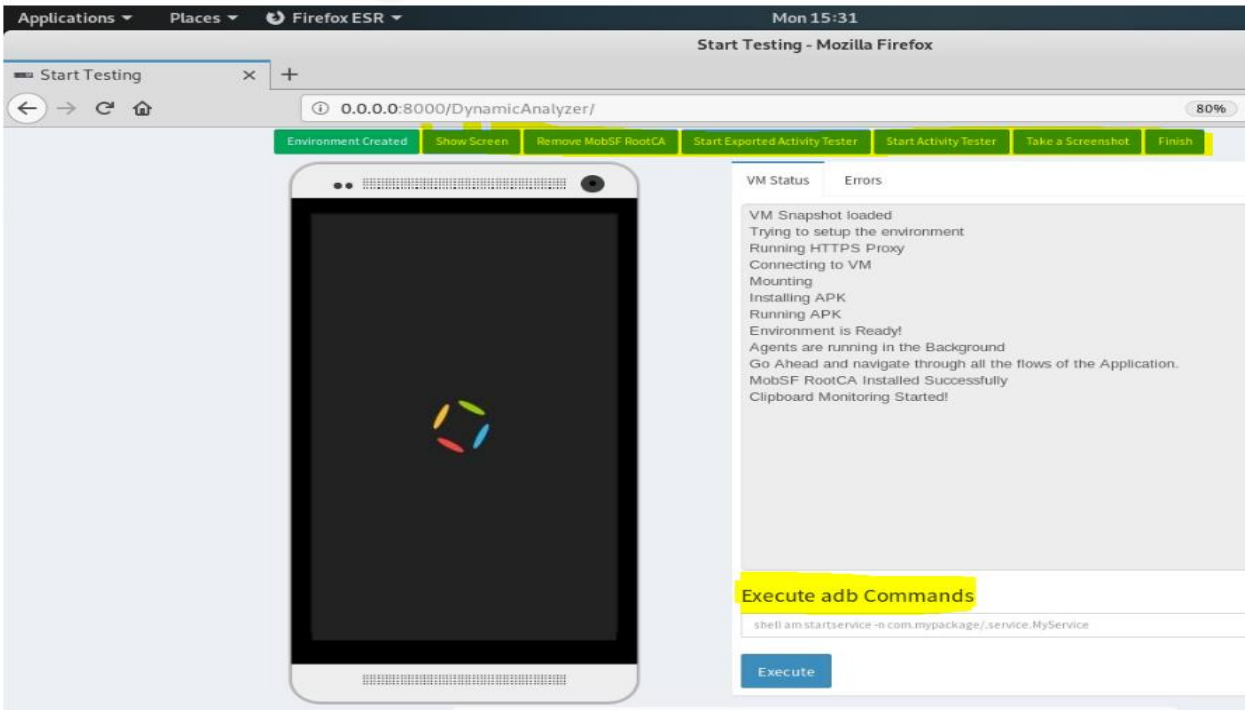


Figure 1. Running ADB commands

MOBSF, which has the buttons marked in Figure 1, such as analysis of activities, taking screenshots, also allows the execution of ADB commands. As a result of these steps, an automatic report is created. The report includes the following results:

- HTTP (s) traffic
- Log analysis
- API analysis
- Application analysis (analysis of data/data/ com.metasploit folder, ie analysis of the data generated by the application)

3.2 Log analysis

The Android log system provides a mechanism for collecting and viewing system debug output. Logs

from various applications and parts of the system are collected in memory, which can then be viewed and filtered with the logcat command.

After establishing a connection to the emulator with ADB Shell, the malicious application was run. From the moment the application was started, the logs started to be seen with the command shown in Figure 2. The obtained logs and analysis results are shown in Table 1 and Table 2.

```
root@dilek:~# adb connect 10.0.2.6:5555
connected to 10.0.2.6:5555
root@dilek:~# adb shell
root@mobsec:/ # logcat v | grep -i "com.metasploit.stage"
```

Figure 2. Log analysis

Table 1. The logs and analysis results obtained with the logcat command 1.

Package Name	"timestamp"	content	Explanation
com.metasploit.stage	1564425028860	\/\/call_log\/calls	An attempt was made to access the phone call history.
com.metasploit.stage	1564425411205	\/contacts	An attempt was made to access the list of people in the phone book.
com.metasploit.stage	1564425669115	\/\/sms\/	An attempt was made to access messages on the phone.

Table 2. The logs and analysis results obtained with the logcat command 2.

Package Name	"timestamp"	method	Explanation
com.metasploit.stage	1564425996140	SMS_SENT	The message has been attempted to be sent.
com.metasploit.stage	1564426783403	getCameraInfo	An attempt was made to access the camera.
com.metasploit.stage	1564427507911	startRecording	It is seen that there is an attempt to record a sound.

3. 3 Network analysiss

When the application was installed on the emulator and started, network registration was made with the wireshark program. When the pcap file created as a result of the recording was analyzed with wireshark, the following results were obtained.

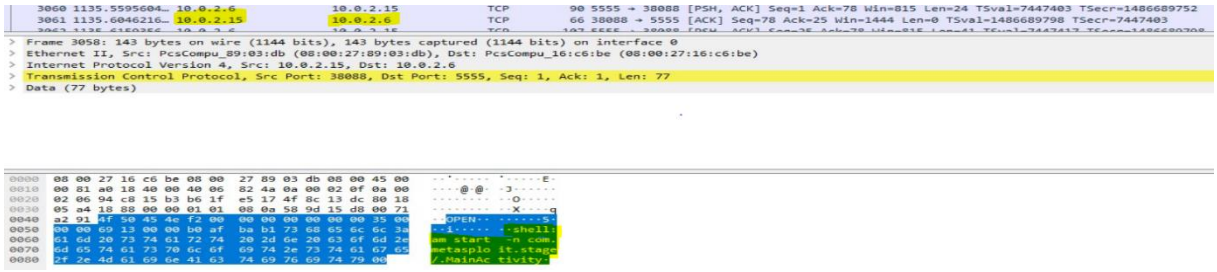


Figure 3. Wireshark network analysis screenshot

When the related file was analyzed, it was seen that the 10.0.2.6 IP address and 10.0.2.15 IP address of the emulator were constantly trying to establish a connection. When the connection activities are examined, it is seen that the malicious application is tried to be started with the start command as shown in Figure 3.

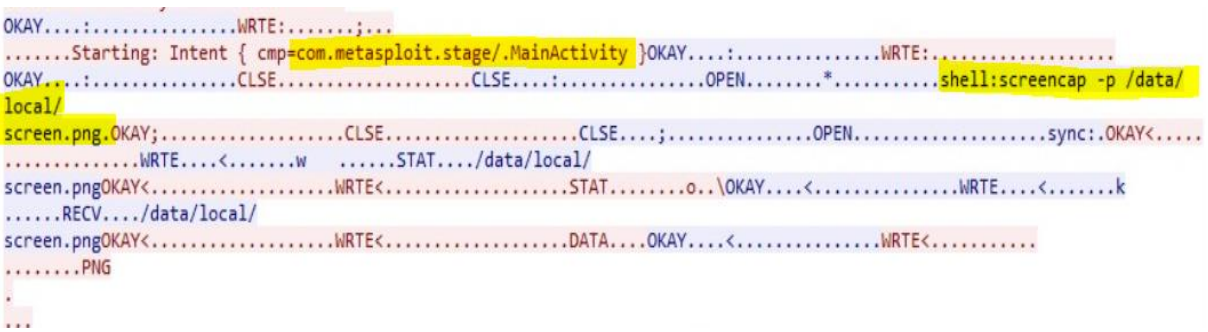


Figure 4. Wireshark network analysis screenshot

When the examination continued, it is seen that the relevant application tries to take a screenshot from the same IP address as shown in Figure 4.

3.4 Process and heap analysis

Today, it is not possible to get a memory dump on mobile devices. Applications on the Android operating system are written in java. Applications written in Java are interpreted with JVM (Java Virtual Machine).

Since the codes interpreted with the JVM cannot be understood by the CPU, the JVM works like a processor. Creates heap space on JVM memory. Object references of the application start to be kept in the static field, and the objects start to work on the heap area.

The processes running on the memory can be listed with the "PS" command to the Android operating system.

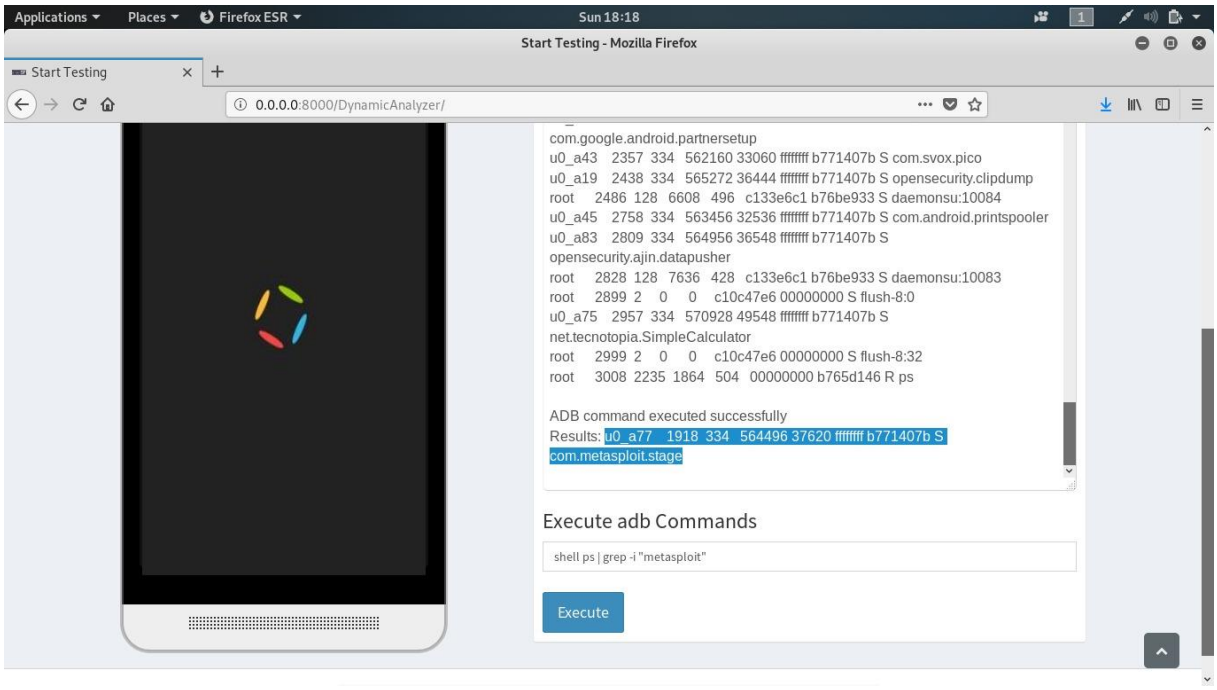


Figure 5. Output of the "PS" command.

When the output of the written command is examined, the information in Table 3 is reached.

Table 3. "PS" command output

Username of the process owner	U0_a77
Process ID (PID)	1918
Parent process ID (PPID)	334
Virtual memory size (VSIZE)	564496
Real memory size (RSS)	37620
Memory address of the event where the process is waiting (WCHAN)	ffffffff
Computer	b771407b
Process name	s com.metasploit.stage

The process named "com.metasploit.stage" of the malicious application runs and when the parent process id is checked, it is understood that it works depending on the root user. It is possible to interfere with the names of processes running with the prctl system call, but the malware name cannot be changed. Because the content of "proc / 1918 / com.metasploit.stage" will not change.

It is possible to dump the heap field with adb. The command is as shown in Figure 6.

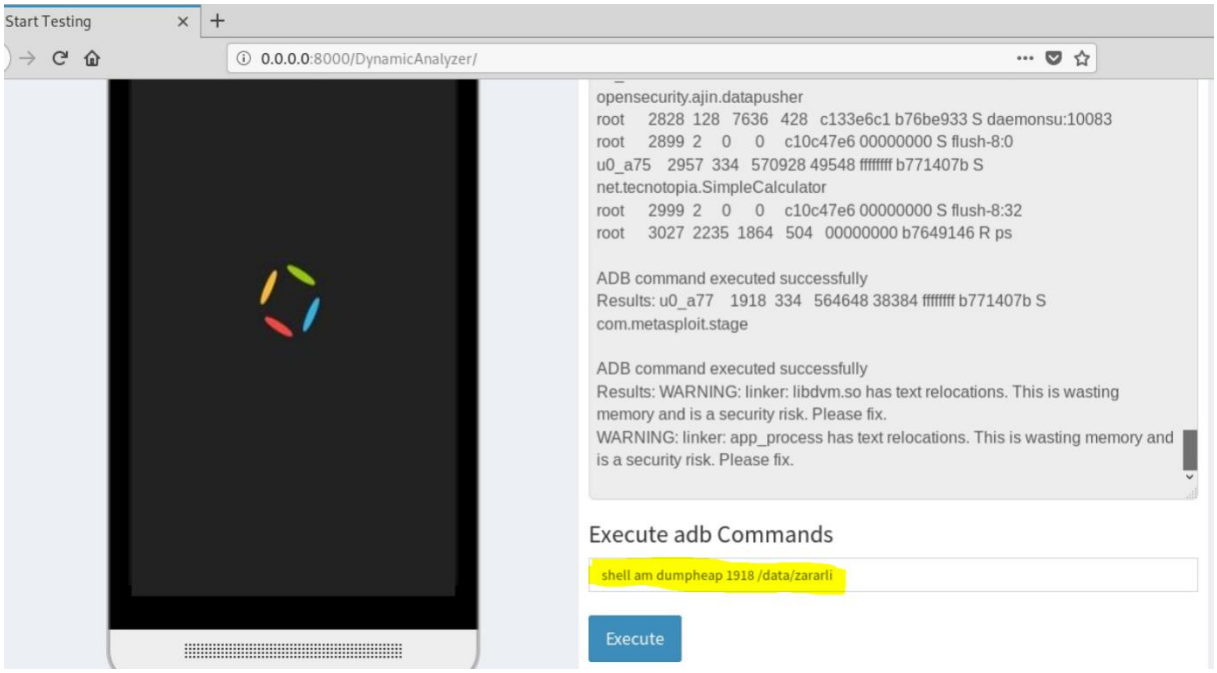


Figure 6. Command to dump heap with adb.

The imported malicious named heap dump must be converted to standard form for analysis. The command required for this is shown below.

```
C:\platform-tools>hprof-conv.exe zararli memory.hprof
```

Figure 7. Command used to convert heap dump to standard form.

Various software and online sites are available for dynamic field investigation. Eclipse Memory Analyzer is used as an application.

The IP address that the application connects to can be detected from the relevant package in Figure 8 below.

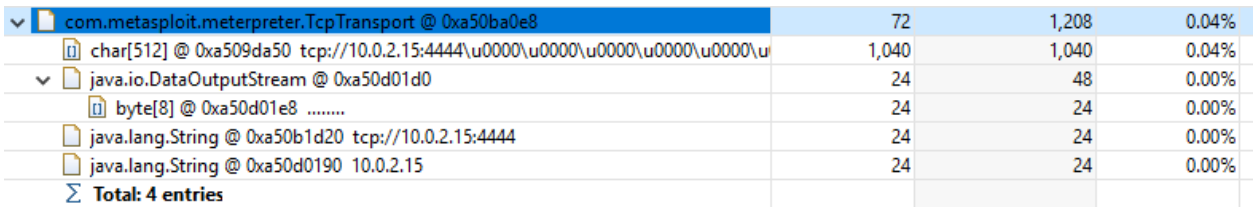


Figure 8. Connected IP address

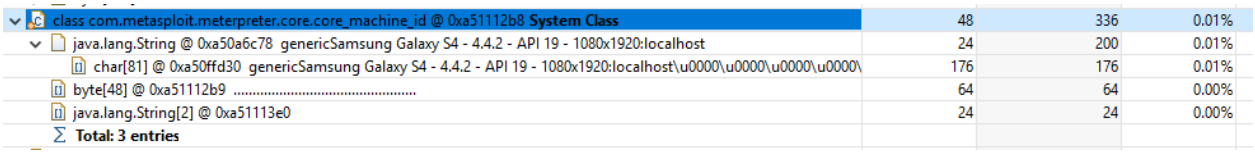


Figure 9. Details of the connected device

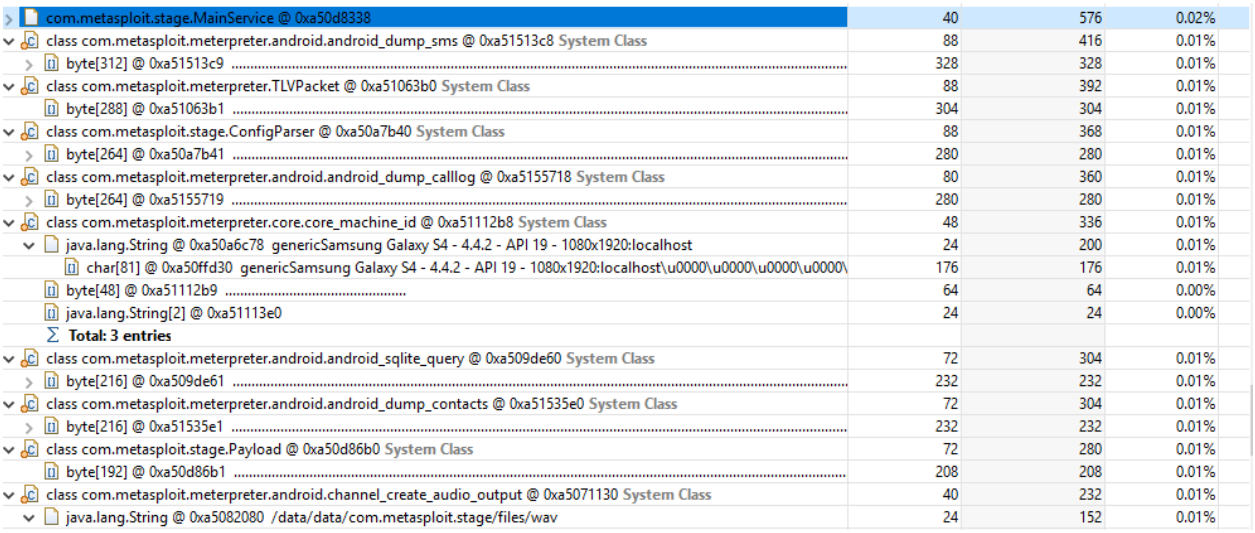


Figure 10. Operations after the connection is established.

4 Result

As a result of the analysis, it is seen that it is very important to make the necessary checks on whether any application that they are downloading should actually access the permissions they want. For example, the application we download to watch movies should only play movies and do nothing else.

Therefore, it is not necessary to request permission to send or receive SMS. One of the biggest disadvantages of Android apps is the inability to install an application on the device without agreeing to grant all the accesses that the apps request.

Here are the steps an Android user can take to avoid malware attacks aimed at gaining access to their device, money and information:

1) People are required to do their own security checks. Before installing a new application, you should read the comments of other users to see if the application is safe and what it demands, what it does. One should be very careful about "free apps" distributed on lesser-known sites or via links sent by email and text message.

2) The permissions requested while installing the application on the device should be read. Make sure that privacy is not invaded by accessing personal information that it does not need, such as your contacts, camera, messages.

3) Installation options should be limited. The official Google Play store or reputable stores such as the Amazon App Store should continue to be used. A quick and easy way to prevent unsafe applications is to make sure that the "install from unknown sources" option remains turned off. This prevents you from installing apps from anywhere other than the Play Store.

4) Application usage should be limited. Only the applications you need should be installed and used regularly. If you are no longer using an application, the application should be uninstalled to prevent unnecessary access to your information. This will help you save memory and reduce your risk of exposure to potential risks.

References

[1] İlker, K., *Truva Atı Zararlı Yazılımlarına Yaklaşım ve Çözüm Önerileri.* Bilgi Yönetimi. **2**(1): p. 28-33.

[2] Çakır, S. and M. Kesler, *Bilgisayar güvenliğini tehdit eden virüsler ve antivirüs yazılımları.* XIV. Akademik Bilişim Konferansı Bildirileri, 2012: p. 551-558.

[3] Data, G., *Mobile malware report.* Retrieved September, 2015. **2**: p. 2015.

[4] Karataş, G., A. Akbulut, and A.H. Zaim, *Mobil cihazlarda güvenlik–tehditler ve temel stratejiler.* 2016.

[5] Van Der Veen, V., H. Bos, and C. Rossow, *Dynamic analysis of android malware.* Internet & Web Technology Master thesis, VU University Amsterdam, 2013.

[6] Utku, A. and İ.A. Doğru, *MOBİL KÖTÜCÜL YAZILIMLAR VE GÜVENLİK ÇÖZÜMLERİ ÜZERİNE BİR İNCELEME.* Gazi Üniversitesi Fen Bilimleri Dergisi Part C: Tasarım ve Teknoloji, 2016. **4**(2): p. 49-64.

[7] Canbek, G. and Ş. Sağiroğlu, *KÖTÜCÜL VE CASUS YAZILIMLAR: KAPSAMLI BİR ARAŞTIRMA.* Journal of the Faculty of Engineering & Architecture of Gazi University, 2007. **22**(1).

[8] Coursen, S., *The future of mobile malware.* Network Security, 2007. **2007**(8): p. 7-11.

[9] DOĞRU, İ.A. and C.A. DİNÇER, *ANDROID KÖTÜCÜL YAZILIM TESPİTİ YAKLAŞIMLARI.* Uluslararası Bilgi Güvenliği Mühendisliği Dergisi. **3**(2): p. 48-58.

[10] Etaher, N. and G. Weir. *Understanding the threat of banking malware*. in *Cyberforensics 2014-International Conference on Cybercrime, Security & Digital Forensics*. 2014.

[11] Ünlü, U., İnternet Bankacılığı Sisteminde Tüketicilerin Karşılaşacağı Olası Saldırılar ve

[12] Çözüm Önerileri. Prof. Dr. Hakan Yıldırım/Buket Alkan, 2018: p. 82.

[13] ÇELİK, S. and B. ÇELİKTAŞ, *GÜNCEL SİBER GÜVENLİK TEHDİTLERİ: FİDYE YAZILIMLAR.* Cyberpolitik Journal, 2017. **2**(4): p. 296-323.

[14] Mercaldo, F., V. Nardone, and A. Santone. Ransomware inside out. in 2016 11th International Conference on Availability, Reliability and Security (ARES). 2016. IEEE.

[15] Kılıç, Ç., *Dünden bugüne fidye yazılımların (Ransomware) gelişimi ve geleceği*. 2019, İstanbul Bilgi Üniversitesi.

[16] Canbek, G. and Ş. Sağiroğlu, *CASUS YAZILIMLAR: BULAŞMA YÖNTEMLERİ VE ÖNLEMLER.* Journal of the Faculty of Engineering & Architecture of Gazi University, 2008. **23**(1).

[17] Uscilowski, B., *Mobile adware and malware analysis.* Symantec Corp, 2013.

[18] **1**.Erturk, E. *A case study in open source software security and privacy: Android adware*. in *World Congress on Internet Security (WorldCIS-2012)*. 2012. IEEE.

[19] Arslan, B., M.S. Gündüz, and Ş. Sağıroğlu. *Güncel Mobil Tehditler ve Alınması Gereken Önlemler*. in *Conference: International Symposium on Digital Forensics and Security*. 2015.

[20] Developers, A., *What is android?* developer. android. com/.../what-is-android. html. Diunduh tanggal, 2011. **14**.

[21] TUFAN, M., et al., AÇIK KAYNAK MOBİL İŞLETİM SİSTEMİ: ANDROID İŞLETİM SİSTEMİ. 2012.

[22] Ehringer, D., *The dalvik virtual machine architecture.* Techn. report (March 2010), 2010. **4**(8).

[23] Liu, J. and J. Yu. *Research on development of android applications*. in *2011 4th International Conference on Intelligent Networks and Intelligent Systems*. 2011. IEEE.

[24] Raja, H.Q., *Android Partitions Explained: boot, system, recovery, data, cache & misc.* Addictivetips. com., May, 2011. **19**.

[25] Bhatia, T. and R. Kaushal. *Malware detection in android based on dynamic analysis*. in *2017 International Conference on Cyber Security And Protection Of Digital Services (Cyber Security)*. 2017. IEEE.

[26] Kapratwar, A., F. Di Troia, and M. Stamp. *Static and dynamic analysis of android malware*. in *ICISSP*. 2017.

Researching and Implementing Secure Data Collection and Data Destruction Methods in Digital Systems

Nur Sena Atalay[1*], Şengül Doğan[1]

[1]*Fırat University, Elazığ, TURKEY*

Abstract

Electronic evidence has become a proven evidence system through computer and computer systems in the Criminal Procedure Board, depending on technological developments. These evidences, which bear the quality of evidence, play a crucial role in the elucidation of information crimes. It is very important and precise to comply with criminal procedure rules and technical rules in order to collect, examine and protect the integrity of the evidence to be examined under criminal procedural laws. Along with the rapid development of communication via social networks due to technology, crime and punishment rates in electronic environment have increased. Due to the inadequacy of conventional evidence collection methods in forensic cases resulting from the abuse of computers and other electronic devices, it has become a necessity to use a special method and techniques for evidence in a case involving computers and electronic devices. In daily life, a lot of data is stored on digital systems. This data may be worthless data, or it can be important data such as personal data, sensitive data of special nature, corporate policies, credit card numbers, financial records. The method of destroying and collection the data is as important as ensuring the security of the data in the storage area. When data needs to be destroyed, it must be destroyed in a safe and permanent manner. When data needs to collect, it must be collected as write blocker. In this study, safe data collection and data destruction methods on digital systems will be researched and applied practically.

Keywords: *Data Collection, Digital Forensics, Data Wiping, Digital Evidence, Digital Forensics Tools*

1 Introduction

Abuse acts such as corruption, theft, unauthorized access, etc. have become widespread in digital systems due to the development of technology. In many organizations such as private sector, banking, public institutions, etc., almost all of the data is stored in electronic systems. Therefore, the recovery and security of electronic data, fraud investigations have a very important role. Digital forensics is the process of collecting, analyzing and reporting electronic data to be accepted as evidence in courts [1]. Digital forensics experts; It provides clarification of all kinds of crimes involving electronic data in incidents related to the fight against corruption, theft, fraud, criminal crimes, financial and banking crimes, internet banking frauds, sexual abuse of children, drugs, etc. Computers are a very important data source for forensic experts, as they contain large amounts of data, keep logs, e-mail transactions, internal hard disk, user accounts and the most widely used electronic devices. Electronic evidence is mostly found on the hard drive. These data are divided into permanent and non-permanent. Non-persistent data disappears when the computer's power switch is turned off. Therefore, volatile information that may be evidence should be kept intact. On the other hand, permanent information is stored on the hard disk if the computer is turned off. During the digital evidence collection phase; the confidentiality, integrity and availability of data must be ensured. An exact copy of the original evidence should be kept and no examination should be made on the original evidence. The process of permanently deleting important data on digital systems is called wipe. Data destruction is just as important as securing the data in the storage space. When data needs to be destroyed, it must be destroyed securely and permanently.

2 Digital Evidence

Digital evidence is data stored on digital systems that can be trusted in court. It can be found on a lot of different systems. For example;

* Corresponding author: nursena.atalay@gmail.com

- Network Tools (Modem, Switch, Router)
- Printers, Scanners and copiers
- Removable Backup Units (Floppy, CD, DVD, USB)
- Credit Card Readers
- Digital Watches
- Digital Cameras
- Mobile Phones, GPS
- PDA and Palm Devices
- External Hard Drives
- Memory Cards
- All Kinds of Log Recorders (Network, OS)
- Computer Systems (desktop, laptop, server)
- Computer Components (HDD, Memory)
- Access Control Tools (Smart cards, biometric scanners)

In addition to the electronic devices listed above, all devices such as credit card duplicating devices, mobile phone cloning devices, GPSs that may be associated with crime are considered as evidence.

2.1. Collection of evidence

The crime scene where the cyber incident took place may include different types of digital evidence. In the system where the cyber attack occurs, multiple digital devices and systems can be interconnected and the attack can affect other systems. There are points to consider when collection evidence [1].

- Before the evidence is collected, the crime scene is documented. Documentation is required throughout the investigation process. These documents should include the device's operating status (open, closed, sleep mode) and physical characteristics such as brand, model, serial number, connections [2].
- Written notes, photographs and / or video recordings of the crime scene and evidence are also needed to document the crime scene and evidence.
- Data must be properly cloned bit to bit. Digital signature (hash) values of the images should be taken.
- The system must be separated from the network to which it is connected. The possibility of data alteration through remote access should be avoided.
- If the computer is on, it should not be turned off, if it is turned off, it should not be turned on.
- The disk to be backed up must be larger than the suspect device's disk size.
- Evidence should be made with a safe copy method, software or hardware. The log file should be saved in the transfer.
- Analysis and investigations should be done on cloned(backup copy) data to preserve the authenticity of the original source.
- If the computer is turned on, volatile data such as Ram should be imaged.
- During the data collection phase, it must be collected as write blocker.

2.2. Destruction of evidence

Evidence destruction is the process of securely deleting sensitive information contained within the storage unit of the digital device. This process can be physical or logical. It is almost impossible to obtain data from the storage device after the data is physically or logically destroyed. Logical destruction overwritten by random bytes, this method is also quite effective. Attempts to collect evidence will be negative if sensitive information is securely deleted. Many internationally accepted standards provide broad definitions for preventing uncontrolled access to data and destroying data. These standards; It is a written information security plan of institutions, controlling and restricting access to data, taking necessary protection measures against the misuse of data. Some of these standards are;

- ISO / IEC 27001
- EU Data Protection Directive
- NISP (National Industrial Security Program)
- Base II Capital Accord
- Data Protection Act
- FACTA (The Fair and Accurate Credit Transactions Act)
- GLB (Gramm-Leach Bliley)

- Sarbanes-Oxley Act (Sox)
- PCIDSS (Payment Card Industry Data Security Standards
- HIPAA (Health Insurance Portability and Accountability)
- NIST (The National Institute of Standards and Technology)

Secure data destruction varies according to the type, size, method of deletion, and the type of media [3].

3 Evidence Collection Tools and Methods

The stage of collecting evidentiary data in digital research varies according to the type of data, stored media, operating system, and size. When categorizing tools and methodologies, they are divided as Windows, Linux, Mac-based and open source. The methods and tools properties used to collect data are given below [4].

3.1 Windows based tools and methods

Crimes that occur in digital environments are called cybercrime. Digital forensics experts manage data identification, collection, protection, analysis and reporting processes in order to detect the committed crime. Some of the tools for Windows-based used to collect evidence are as follows.

3.1.1 Accessdata FTK imager

Data collection is one of the most important steps in digital investigation. One of the most reliable methods in the data collection phase is to take the image of the storage unit of evidence while preserving the confidentiality and integrity of the data. FTK Imager is one of the software used to take images. It was developed by the AccessData company and is free of charge. Hard drives, USB memory cards, Zip drives, CDs and DVDs, directories or files can be imaged with FTK. At the same time, it mounts the relevant disk image files as Read-only, prepares an environment for these images to be treated as if they are a hard disk drive through Windows Explorer, and allows the directory or files to be copied out of the image files. There are two versions as "FTK Imager" and "FTK Imager Lite" for use in Windows environment. FTK Imager requires installation, but FTK Imager Lite does not require any installation.

There are also versions that will work on the operating systems of Ubuntu, Fedora and Mac systems. The integrity of the data is preserved in the images taken with the FTK Imager. Hashing algorithms such as MDH5, SHA1 are used to protect data integrity. However, it can generate hash reports for normal files and disk images, and then these hash values can be used for integrity verification. FTK Imager allows you to create disk images in four different formats. These are raw (dd) format, SMART format, E01 format and AFF format. When starting the image processing, source evidence type can select as Physical Drive, Logical Drive, Image File, Contents of a Folder and Femico Device(multiple CD/DVD) [5]. The physical image type is selected in the data storage unit where all volumes are examined. The logical image acquisition method is selected in the data storage unit where a particular partition / file will be examined.

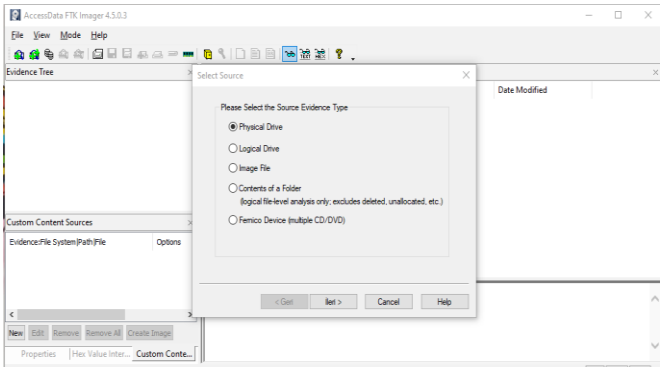


Figure 12. Source evidence type

Custom Content Image option can be selected to securely transfer the evidential a file or files.

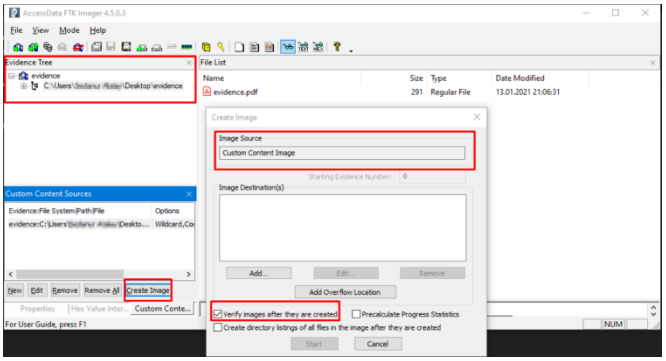


Figure 13. Custom Content Image

During the data collection stage, information such as Case Number, Evidence Number, Unique Description, Examiner, Notes can be defined.

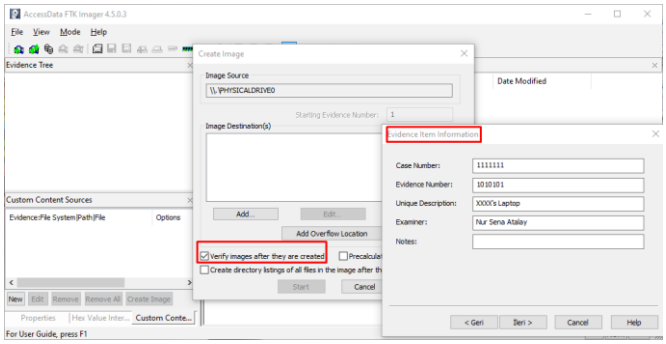


Figure 14. Information of investigation in data collection stage

3.1.2 Robocopy

Robocopy is provided to the user as a default feature of the Windows operating system. With Robocopy, data is copied securely and hash values calculated. The integrity of the data is ensured with the hash value. Robocopy is run on the Windows operating system via the terminal screen. This command line based tool is located at C: \ Windows \ System32 \ Robocopy.exe by default. The Robocopy tool provides different copying possibilities with various command combinations [6]. Also log file can be created regarding the copying process. Some of these cominations are;

- Robocopy C:\Data D:\Data /Move
- Robocopy C:\Data D:\Data /MIR
- Robocopy C:\Data D:\Data /E
- Robocopy C:\Data D:\Data /E /S /XO > C:\robocopylog.txt

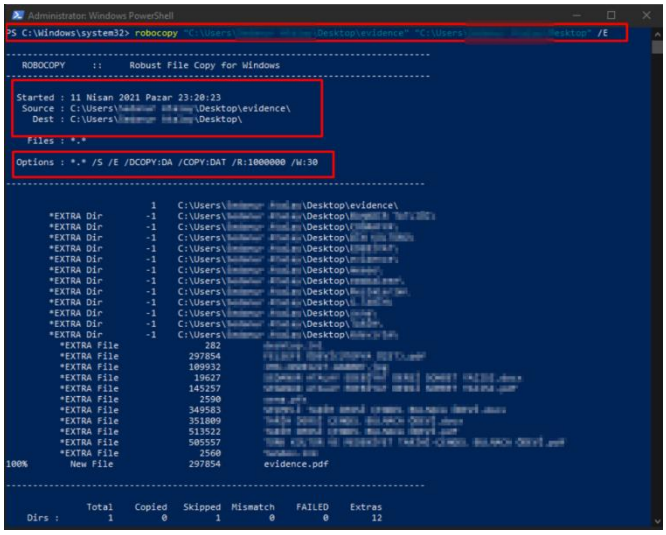


Figure 15. Robocopy using

3.1.3 X-Ways forensics

X-Ways Forensics is a licensed forensics investigation product created by forensics experts. Image acquisition and analysis operations can be performed with X-Ways software. It is designed to run on 32/64 Bit platform of Windows operating systems [7]. The features that are provided are as below:

- Disk cloning, imaging and viewing
- View image files in dd, VHD and VMDK format
- CRC32, MD5, SHA-1, SHA-256 algorithm signature (hash) calculation for files
- Ability to show existing and deleted files in all subdirectories
- Support E01 image files
- Write protection feature not to spoil the originality of the data
- Passwors recovery.

3.1.4 Encase forensics imager

EnCase is a digital data review and analysis program created by a company called Guidance Software in California, United States. It is this program that proves itself in terms of evidence analysis and evidence association. It is commonly used by many law enforcement organizations around the world, especially the FBI. The reports and data submitted to the court by EnCase have been accepted as evidence. EnCase evidence files and EnCase logical evidence files can be created with EnCase Forensic Imager. This tool, whose user interface is very similar to EnCase Forensic software, does not have evidence processing, preview and analysis features. No EnCase license is required for the completely free EnCase Forensic Imager.

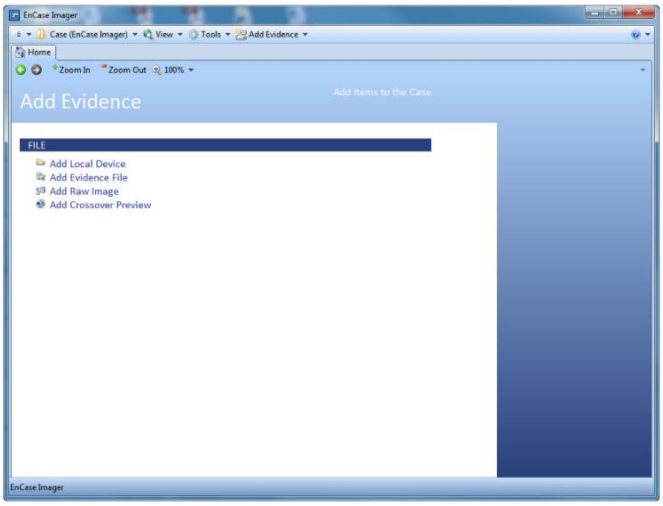


Figure 16. EnCase Forensic Imager

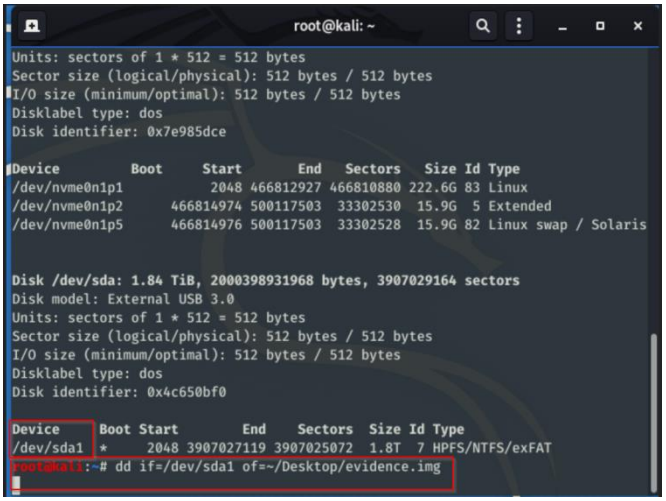


Figure 17. Linux dd image command line

3.2 Linux based tools and methods

Evidence of crimes in the digital environment can be Windows-based or have a Linux-based operating system. Some of the tools for Linux-based used to collect evidence are as follows [8].

3.2.1 The sleuth kit (TSK)

The Sleuth kit (TSK) is a collection of Unix-based command line analysis tools. Autopsy is a graphical interface for TSK.

- Analyzing FAT, NTFS, Ext2/3 file system
- List files and directories
- Recover deleted files
- Generate file activity timelines
- Perform keyword searches

3.2.2. Linux DD (raw) image

DD image is a secure method provided to collect data. With DD imager, it can take images of a hard disk, usb memory, cd / dvd. DD imager is run on the command line in linux system. The DD command takes two basic parameters. Other parameters are optional. -if indicates the source from which the copy will be taken, and -of indicates the destination where the data will be saved. The extension of image files is usually .dd, .001 or .img. The three most commonly used applications to create disk images in raw format are listed below [9].

- dd (Native * nix version)
- dd.exe (Windows version)
- dc3dd

3.2.3 Computer aided investigative environment (CAINE)

CAINE is a Linux live distribution. It aims to provide a collection of GUI and forensic tools. It includes open source tools that support data collection, acquisition, analysis, and analysis stages. It also supports the researcher by providing capabilities to automate the creation of the final report.

4 Evidence Destruction Tools and Methods

Secure data deletion is the case of logical and physical deletion of data in the storage unit of the device without physically damaging the device. Data is always available even in case of formatting a hard disk. In order for the data to be destroyed, it must be deleted by a special method. There are special hardware and software for data destruction. Secure data deletion hardware has been developed by taking different types of drivers into consideration. These hardwares replace the data on all sectors with meaningless data at bit level with post-processing test and reporting feature. In the secure data deletion method, data is written to all sectors on the drive many times with different patterns, so the magnetic signatures of the data are effectively destroyed. Media can be reused after secure data deletion. Secure data deletion varies according to the type of media.

Table 6. Destruction capabilities per disc type

Media Type	Safe Destruction	Physical Destruction
Floppy Disk	Yes	Yes
Hard Disk	Yes	Yes
USB Hard Disk	Yes	Yes
Zip Disk	No	Yes
Tapes	No	Yes
CD / DVD	No	Yes
Memory Cards	Yes	Yes
USB Flash Disk	Yes	Yes
BlueRay Optics	No	Yes

The most used wipe methods are as follows,

- Gutmann is the process of writing data 35 times on the hard disk using the Gutmann algorithm. This process is the most reliable wipe method. However, the deletion process takes too long.
- Schneier is Bruce Schneier's algorithm. It wipes the file 7 times in a row with a hidden random number generator. It is a very safe deletion method.
- Dod-7, this method is the US Department of Defense file deletion system Dod 5220.22-M ECE. Wipes a file by overwriting it 7 times.
- 5 Pass Gutmann, based on the Guttman 35 Pass, this method shuffles the file you want to destroy by overwriting it 5 times.
- 3 Pass DoD3, overwrites a string of numbers 3 times and is more convenient to use only for non-private files. It is preferred because of its short processing time.
- 3 Pass AFS SI5020, outputs the document to be deleted exactly three times. Because it uses the 3 pass method, it creates a medium level security level like the others.
- 1 Pass Random: This method is not safe at all and is very simple. Before deleting a file, it processes random information on it and is not that difficult to retrieve. .

4.1 Data destruction softwares

In order for the data to be destroyed, it must be deleted by a special method. There are special hardware and software for data destruction.

4.1.1 Ccleaner

CCleaner is used as a system cleaner to remove Windows files such as file, internet browser, cache, etc. It also includes a tool that can completely destroy all data on a drive. CCleaner supports both SSDs and mechanical drives. Data destruction methods are as follows:

- DoD 5220.22-M
- Gutmann
- Schneier
- Write Zero

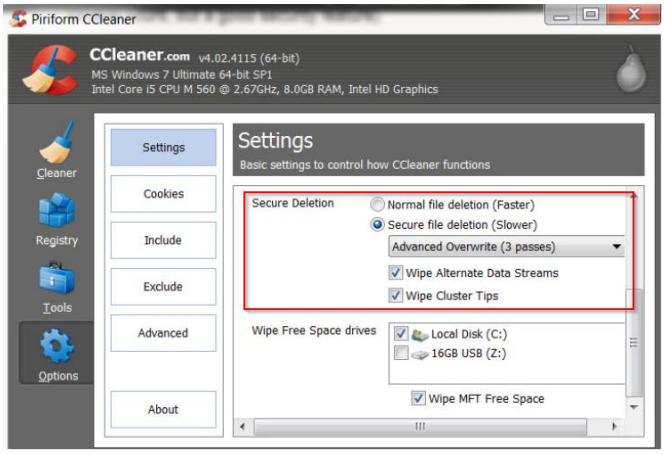


Figure 18. Ccleaner wipe methods

4.1.2 SDelete

SDelete is a command line based data destruction utility. It is a Windows based tool and can be operated from the Windows command screen. It uses DoD 5220.22-M method as data destruction. The SDelete tool is a free tool available from Microsoft [10]. It is part of the Sysinternals Suite.

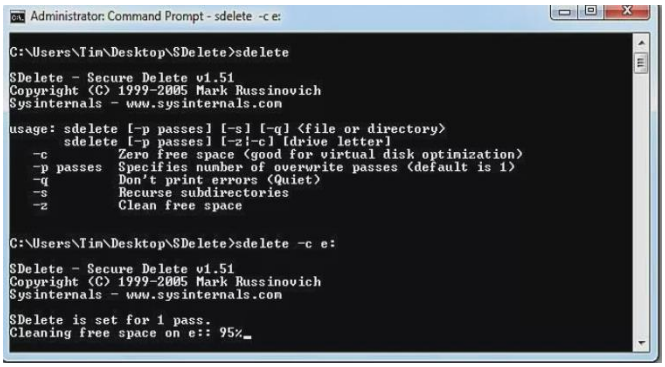


Figure 19. Sdelete command line using

5 Conclusion

In this study, safe data collection and data destruction methods on digital systems have been researched and applied practically. It has been dwelt on what tools and method used to collect or destroy evidence in a digital forensics investigation. All the tool and method used have different characteristics and used in different operating system.

6 Acknowledgment

I am very much thankful to Doç. Dr. Şengül DOĞAN for their guidance throughout this paper.

References

[1] B. V. Prasanthi "Cyber Forensic Tools: A Review". *International Journal of Engineering Trends and Technology (IJETT),* Volume-41 Number-5 - November 2016

[2] Nana Rachmana Syambas and Naufal El Farisi. "Development of Digital Evidence Collection Methods in Case of Digital Forensic Using Two Step Inject Methods", *School of Electrical Engineering and Informatics, Bandung Institute of Technology*, October 2014

[3] Miroslav Ölvecký, Darja Gabriška. "Wiping techniques and anti-forensics methods" *IEEE 16th International Symposium on Intelligent Systems and Informatics* September 13-15, 2018, Subotica, Serbia

[4] Kambiz Ghazinour . "A study on digital forensic tools". *IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)* 2017

[5] Yvonne LeClaire. "The Forensic Process Examined: Creating cases for classroom use*", Lewis University* MSIS 68-595

[6] RoboCopy. [Online]. Available:*https://docs.microsoft.com/en-us/windows-server/administration/windows commands/robocopy,* Accessed: April 2021.

[7] Brett Shavers,"X-Ways Forensics and Electronic Discovery", *In book: X-Ways Forensics Practitioner's Guide (pp.197-209*), December 2014

[8] S A Finney, "Real-time data collection in Linux: a case study", 2001 May;33(2):167-73

[9] Moses Ashawa, Morris Ntonja, "Design and Implementation of Linux based Workflow for Digital Forensics Investigation", *International Journal of Computer Applications (0975 – 8887) Volume 181 – No. 49,* April 2019

[10] Michael Freeman, Andrew Woodward, "*Secure State Deletion: Testing the efficacy and integrity of secure deletion tools on Solid State Drives*" International Journal of Computer Applications (0975 – 8887) Volume 181 – No. 49, April 2019