MACHINE LEARNING MODELS FOR NETWORK INTRUSION DETECTION AND AUTHENTICATION OF SMART PHONE USERS

A Thesis

Presented to

the Faculty of the Elmer R. Smith College of Business and Technology

Morehead State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

---

by

S. Sareh Ahmadi

November 18, 2019

ProQuest Number: 27664742

All rights reserved

INFORMATION TO ALL USERS
The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript
and there are missing pages, these will be noted. Also, if material had to be removed,
a note will indicate the deletion.

ProQuest 27664742

Published by ProQuest LLC (2019). Copyright of the Dissertation is held by the Author.

All rights reserved.
This work is protected against unauthorized copying under Title 17, United States Code
Microform Edition © ProQuest LLC.

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

Accepted by the faculty of the Elmer R. Smith College of Business and Technology, Morehead State University, in partial fulfillment of the requirements for the Master of Science degree.

_____
Dr. Sherif Rashad
Director of Thesis

Master's Committee: _____, Chair
Dr. Sherif Rashad

_____
Dr. Ahmad Zargari

_____
Dr. Heba Elgazzar

_____
Date

MACHINE LEARNING MODELS FOR NETWORK INTRUSION DETECTION AND AUTHENTICATION OF SMART PHONE USERS

S. Sareh Ahmadi
Morehead State University, 2019

Director of Thesis: _____
Dr. Sherif Rashad

Machine learning techniques have been utilized in many areas of security such as computer networks security and smart phone user authentication due to their unique properties such as their ability to automatically learn and improve from experience, adapt quickly to new and unknown challenges, and high accuracy. For these reasons, this study utilizes machine learning methods for building efficient intelligent models for intrusion detection systems (IDSs) in computer networks, and for smartphone user authentication based on performing daily living activities.

Network security consists of protection of access, misuse, and monitor unauthorized access in a computer network system. Network security systems consists of a fire wall, antivirus and an intrusion detection system (IDS). The IDS monitors a network traffic to find suspicious activity, such as an attack or illegal activities. Many researches have focused on different machine learning methods to improve the performance of IDS. Due to availability of irrelevant or redundant features or big dimensionality of dataset, which results in inefficient detection

process, this study focused on identifying important attributes in order to build an effective IDS. A majority vote system, using three standard feature selection methods, Correlation-based feature selection, Information Gain, and Chi-square is proposed to select the most relevant features for IDS. The decision tree classifier is applied on reduced feature sets to build an intrusion detection system. The results show that selected reduced attributes from the proposed feature selection system give a better performance for building a computationally efficient IDS system.

User authentication is one of the important problems in smart phone security. Technological advancements have made smartphones to provide wide range of applications that enable users to perform many of their tasks easily and conveniently, anytime and anywhere. Many users tend to store their private data in their smart phones. Since conventional methods for security of smartphones, such as passwords, PINs and pattern locks are prone to many attacks, this thesis proposes a novel method for authenticating smartphone users based on performing seven different daily physical activities and extracting behavioral biometrics, using smartphone embedded sensor data. The proposed authentication scheme builds a machine learning model which recognizes users by the proposed method. Experimental results demonstrate the effectiveness of the proposed framework.

Accepted by: _____, Chair
Dr. Sherif Rashad

_____
Dr. Ahmad Zargari

_____
Dr. Heba Elgazzar

Table of Contents

CHAPTER 1: INTRODUCTION

The first section of this chapter provides a background related to intrusion detection systems and their importance for security of the computer networks**.** The second part of this chapter, provides a background related to smartphone security with a focus on the problem of user authentication.

1.1 Network Intrusion Detection

In today's world, computers and computer networks connected to the internet play a major role in communications and information transfer. In the meanwhile, profitable individuals have taken action against the computer systems to get access to important information of special centers or other people's information with the intention of imposing pressure or even disruption of the order of systems. Therefore, the need to maintain information security and maintain efficiency in computer networks that are connected with the outside world is completely tangible. An intrusion detection system (IDS) can be a set of tools, methods, and documentation needed to identify, and report unauthorized network activities (Buczak, 2016). In fact, intrusion detection systems monitor activities in the network, by using algorithms that identify suspicious activities and introduce them as intrusion. However, it is common that some of these activities that are not intrusive are still detected as intrusion incorrectly. This is the reason that so many research efforts have been dedicated to improve the performance of intrusion detection systems.

There are two approaches for detection of intrusion: misuse (or signature detection) and anomaly detection. The first one uses the known attack patterns and signature that have already been recognized, while the second technique compares the deviation from the normal behavior of the monitored network devices (Boujnouni, 2018). Misuse detection can identify malicious activity in the network without high false alarm but it is only capable of detecting known attacks. On the other hand, anomaly detection can detect both known or unknown

attacks. This is very important feature since networks are constantly subject to new kinds of intrusions. One of the methods of anomaly detection is based on using machine learning and data mining algorithms to learn from a training dataset and construct a model to classify network activities as normal or attack. One of the bench mark datasets in network security is KDD CUP'99. However, a study on the this dataset (Tavallaee, 2009) shows that, there are some drawbacks in this data set. A statistical analysis was conducted, and deficiencies are found out for KDD CUP'99. The NSL-KDD dataset was suggested to solve the problems of KDD CUP'99 (Tavallaee, 2009). According to their study, KDD CUP'99 was full of redundant and duplicate records which result in a biased machine learning model toward the frequent records. In the refined version of this dataset they removed all repeated records. Moreover, the new dataset contains reasonable number of records which means any experiment can be done on the whole data set without randomly selecting a sample. Evaluating methods such as accuracy, detection rate and false positive rate on the KDD dataset is not an appropriate option. To solve this problem, the number of selected records from each difficulty level group in the new version is inversely proportional to the percentage of records in the KDD dataset.

This thesis, introduces a novel method for finding the most relevant features that can contribute to build an efficient machine learning model for detecting attacks in intrusion detection systems.

1.1.1 Research Goals and Objectives for Network Intrusion Detection

This research focuses on application of machine learning methods in networks security for the problem of anomaly-based network intrusion detection to decide whether or not an intrusion is taking place on a network. Due to availability of irrelevant or redundant features or big dimensionality of dataset which results in an inefficient detection process, this research work, aims to identify important features for IDS that is computationally efficient and effective for design, implementation and testing machine learning algorithm for intrusion detection system.

There are three main objectives for the creation of an IDS

- Proposing a feature selection technique for the datasets to reduce the complexity of the IDS and improve classification accuracy
- Applying machine learning models for IDS based on the selected features to effectively predict intrusions.
- Testing the developed algorithms on a real world datasets

1.2 Smart Phone Authentication

Smart phones have become increasingly more popular these days due to their applications in human's life for performing different tasks such as bank transactions, paying for public transports, accessing social media accounts, receiving and sending emails and so on. As innovations in smartphone applications are growing rapidly, many companies are encouraged to provide their services through these smartphones as well. As a result, there is a great tendency for all the people of the world to have smart phones. Due to these pervasive purposes and ease of use, many users store their private data in their smartphones. Therefore, smartphone security is becoming increasingly important as people save more sensitive information on their smartphones. The most common approaches for securing mobile phones are PINs, password, pattern lock and finger print scans and face recognitions. However, each of these traditional approaches have their own weaknesses. They are vulnerable to different attacks such as smudge attack which is basically getting oils from users' skin for patterns or PINs detection, or shoulder surfing attack, which are observation techniques such as glancing over the shoulder of a user to obtain information. Passwords and PINs can also be stolen by monitoring users over a period of time (Alzubaidi, 2016). Fingerprint scans are subject to spoofing and additional hardware are needed for them to operate (Ehatisham-ul-Haq, 2018). Face recognition schemes are constantly affected by environmental condition such as light as well (Ehatisham-ul-Haq, 2018). Moreover, these frequently used methods are one-time authentication methods which means

they are not able to authorize a user after the first entry. Hence, they cannot recognize and authenticate smartphone users continuously (Centeno, 2017). Therefore, a continues authentication scheme is essential for security of smartphones. Continues authentication, also referred as implicit, passive or progressive authentication, constantly re-authenticate the individuals when the user is using the smartphone without requiring any specific action from the user (Centeno, 2017) . To address these challenges in user authentication for smartphones, many researchers started using behavioral biometrics authentication schemes which utilizes the embedded sensors in mobile phones (Alzubaidi, 2016). With these techniques, authentication is done by identifying the behavioral traits of smartphone users while they are interacting with smartphones (Alzubaidi, 2016). Most of the behavioral and physiological biometrics are based on built-in sensors which are capable of measuring the motion, position and environment of a device environment. For this reason, this research introduces a scheme that authenticate smartphone users continuously based on performing physical activities as behavioral biometric, using smartphone embedded sensors.

1.2.1 Research Goals and Objectives for Authentication of Smart Phone Users

This research studies the application of machine learning methods in activity recognition and smart phone authentication. Due to limitations of traditional methods for security of smart phones, this research, aims at building a continues authentication schemes that utilizes behavioral biometrics for authenticating smart phone users.

The main objectives for the smartphone authentication

- Proposing a smartphone authentication scheme that is continues and utilizes physical activities as behavioral biometrics.
- Proposing a feature selection technique for finding the most relevant features for building an efficient machine learning model for activity recognition and smart phone user authentication.

- Applying machine learning models for activity recognition and smart phone authentication based on the selected features to effectively predict activities and smart phone users.
- Testing the developed algorithms on a real world datasets

CHAPTER II: LITERATURE REVIEW

The first section of this chapter provides a review of different techniques for anomaly detection for computer networks and the applications of machine learning methods for intrusion detection systems. The second section provide a review of different types of behavioral biometrics and their application in smartphone authentications. Some researches on smartphone authentication based on behavioral biometrics are discussed as well.

2.1. Applications of Machine Learning in Intrusion Detection Systems (IDS)

There are different techniques for anomaly detection. Threshold detection, rule-based measures, statistical measures and machine learning methods. The first technique counts some attributes of user and system behavior and then it compares them with a tolerance level. The second approach tries to define a set of rules that can be used to decide whether a given behavior is normal or not. Statistical measures analyze the distribution of the network features. The last technique is based on machine learning and data mining and it learns from a set of training data and constructs a model able to classify new network traffics as legitimate or malicious. There are various researches on intrusion detection using machine learning methods. The application of various data mining techniques for intrusion detection systems for development of secure information system was discussed in detail in (Wankhade, 2013) and approves that normal behavior inside the data can be understood by that machine learning methods and this knowledge can be utilized for detecting unknown and unnormal behaviors. One of the example of applications of machine learning classification on NSL-KDD was introduced by (Panda, 2010) where a discriminative multinomial Naïve Bayes classifier is applied in order to build a very accurate network intrusion detection system by making the use of filtering analysis. In another study (Boujnouni, 2018), a new version of support vector machines (SVM) was presented for an IDS. The experimental results show that the proposed method has high novelty detection rate of unknown network behavior. An application of

clustering technique for IDS was introduced (Li , 2011) where the k-mean clustering was used with particle swarm optimization (PSO) to have an optimal IDS. Another optimal IDS was presented in (Tao, 2004) where a one-class classification based on support vector domain description (SVDD) with genetic algorithm was proposed. To improve false alarm rate, a novel hybrid intelligent decision method was presented which uses both clustering and classification techniques for attack detection (Panda, 2012). The study in (Panda, 2009) used data mining approach to derive association rules where the knowledge of experts are converted to rules so that a predictive model can be constructed for IDS. The research proposed a method to overcome the complexity of association rules which come from large number of rules. In another research, they divided the NSL-KDD dataset into four category of attributes (basic, content, traffic and host) and then attributes of KDD data set were categorized and formed by all combinations of four classes. A random tree algorithm was applied to raise the suitability of the data set with minimum possible false alarm rate (FAR) (Aggarwal, 2015). A deep learning based intrusion detection system was introduced in (Whang, 2018) in order to prevent an adversary cause model to learn an incorrect decision-making function such as avoiding detection of attacks or classifying benign input to as attack input. The roles of individual features in generating adversarial examples were also explored and reported.

All of the above-mentioned methods proved the effectiveness of machine learning methods for intrusion detection systems, however, they are based on complex computational models due to applying all features in NSL-KDD. For any machine learning method, feature reduction is an important step before building a model for IDS. A number of approaches have been proposed to make the model as efficient as possible. In (Ganapathy, 2015), a new feature selection algorithm was proposed by using an attribute selection and tuple selection which uses rules and information gain ratio for feature selection. They applied the method on KDD dataset which has some drawbacks. Recently some researches have focused on feature selection for

NSL-KDD dataset. Examples include the research (Mukherjeea, 2012) in which a model for feature selection on the basis of feature's vitality was proposed. The vitality of a feature is determined by considering three main performance criteria, the classification algorithm and setting a threshold. Then sequences of searches are performed on different feature sets. The search begins by a set of all attributes on NSL-KDD dataset and removing one feature and checking the metrics to see if they meet the threshold. This process continues to reach the desirable performance. This method improved the results for intrusion detection, however it has complexity and overheads. In another study (Chae, 2013), a feature selection method was proposed and compared with other techniques. The proposed method is based on using attribute ratio that calculates the feature average of total and each classis. A higher accuracy was reported in comparison with other techniques. However, this method was only applied on nominal features and calculation time is required for this method and other methods. A new feature selection method was introduced using correlation feature selection measure (Chang's method) in (Nguyen, 2010) to reduce the dimensionality of the features to provide an optimal subset of features. In this research optimization method was applied to have a new search strategy for obtaining relevant features to make the IDS more efficient but optimization techniques lead to computationally complex method. A new hybrid algorithm PCANNA that combines the conventional principal component analysis (PCA) with neural network algorithm was introduced to reduce the number of attributes on NSL-KDD data set (lakhina1, 2010). However, neural networks make the algorithm computationally expensive. Another study (Kumar, 2016) proposed an updated version of Naive Bayes (NB) classifiers and applied various feature selection techniques for feature selection to see which features contribute most for having the highest accuracy for the novel proposed Naïve Bayes classifier. The gain ratio plus ranker method selects the best features for the novel naïve Bayes classifier in this method, however this feature selection method was not tested on other classifiers and the traditional

Naïve Bayes classifier. The results were for a specific classifier and it is not a general method. In (Assi1, 2017) five classification methods with three feature selection strategies on NSL-KDD dataset were investigated. Each method of attribute selection was applied separately for building each classifier and calculates the performance separately but leading to a time-consuming process. The highest accuracy in (Assi1, 2017) comes from the J48 classifier with information gain feature selection.

Overall, there are a few number of research works that apply feature selection on NSL-KDD dataset. For this reason we propose a novel method for building an effective intrusion detection system by introducing a novel method that selects the most relevant subsets of features in NSL-KDD for an efficient IDS as will be explained in chapter III.

2.2 Applications of Machine Learning in Smart Phone Authentication

According to (Sitova, 2016), a biometric determines the unique physical or behavioral traits of people and tries to identify users correctly. There are two categories for biometrics: behavioral and physiological. Physiological security aims to detect physical characteristics of a user such as retina or iris scans fingerprints, face recognition, finger and palm print (Ehatisham-ul-Haq, 2017). In contrast, the aim of behavioral authentication schemes is learning the characteristics of the behaviors that that is constant for a period of time. They consist of hand movement and waving patterns, keystroke, touch screen interactions, gait patterns, voice, signatures, behavior profiling and activity recognition (Sitova, 2016). Physical biometrics authentications usually require more hardware, as a result, behavioral biometrics are cheaper than physiological biometrics (Ehatisham-ul-Haq, 2017). For this reason, several researches have been dedicated to applying behavioral biometrics for smart phone user authentication. In (Yang, 2015), they discovered the hand waving of different users are unique and they utilized this behavioral biometric for locking and unlocking. Another study introduced an approach based on waving gestures to protect smartphones from harmful attacks by dialing behavior

(Yang, 2015). Authenticating with the nature of typing motion is an old method in which typing motions or keystrokes is used to identify users (Sitova, 2016). A mixed approach based on keystroke and handwriting was proposed and evaluated with a significant accuracy by applying various classification methods (Trojahn, 2013). The authors in (Zheng, 2014) analyzed how a user touches the phone as tapping behavior and a non-intrusive behavioral authentication approach was proposed. A multi-touch gesture-based authentication technique was introduced by classifying the gestural inputs movement characteristics of the center of the palm and fingertips on the multitouch surface of devices (Sae-Bae, 2012). The touch movement of users during pattern input was verified as a biometric behavioral to develop a security method for smart phones (Meng, 2016). A study (Neverova, 2016) showed that human biometrics, have important information about user identity and can serve as a valuable source of authentication systems. As mentioned before, signature behavior is considered as a behavioral biometric. A method based on online signature that is drawn by a fingertip on a mobile device was developed to authenticate people (Sae-Bae, 2014). Another example of behavioral biometric is voice which is used for identification based on recognizing manner and pattern of speaking (Sitova, 2016). To apply this behavioral biometric, a method to identify a speaker who is on the phone call was introduced for user detection (Kunz, 2011). Behavioral profiling is one of the behavioral biometrics that is used for user identification by monitoring how a user interact with digital services and applications. It is divided into two categories, the network base and the host base. The first method monitors behaviors to service providers while the latter investigates where and when users' use different applications (Sitova , 2016). A study on behavioral profiling used a host and cloud approach for user notification about applications that behave badly (Papamartzivanos, 2014). A new approach for user validation is gait biometric. Its purpose is to identify people' walking styles so that verify users based on a person's movement. Gait patterns are introduced, as a promising biometric for recognizing human identities by

acceleration signals using wearable or portable smart devices (Zhang, 2015). In another study they implemented a technique for extracting gait cycle using a function called Gaussian Dynamic Time Wrap (GDTW) to build a similarity measure for classification (Muaaz, 2013). Other approaches for identifying smartphone users is based on using inertial sensors to get the behavioral characteristics of users performing different activities. Study in (Alzubaidi, 2016) summarized the limitations of behavioral biometric approaches for smartphone user authentication for hand waving patters and gestures, keystroke dynamics, touch screen interactions, signature, voice, gait patterns and behavioral profiling. For example, for gait patterns biometrics, the patterns of a user changes by using different outfit, also hand waving and gesture pattern may be the same for multiple users. Behavioral profiling of a user can vary according to their mood such as being sad, happy or exited. Moreover, learning the hand movement and waving patterns for new users is highly time-consuming. keystroke and touch screen biometric, requires active interaction with the touchscreen. The voice is significantly affected by the noise around the users. However, the physical activity recognition as a behavioral biometric for authentication of smartphone can be a reliable biometric source for authenticating users since they are daily living activities and generally people perform these activities multiple times of a day. Recent researchers in security of smart phones have made use of this biometric behavior for the authentication of people. In (Ehatisham-ul-Haq, 2017) introduced an authentication system was proposed based on activity recognition for different classifiers and it was concluded that using Bayes Net classifier is the best option in terms of accuracy and the time needed to recognize the activity. However, any strategy for feature selection was not proposed. In (Ehatisham-ul-Haq, 2018) an authentication schemes based on behavioral traits by using physical activity patterns of different smartphone users was proposed to provide different level of access to users' smartphones. However, different models for six different activities for five different body position were built. As a result they were thirty

different models for authentication. The research in (Ehatisham-ul-Haq, 2017) proposed a probabilistic scoring model for recognizing the activities and incorporated it with user authentication scheme. KNN clustering technique was applied for selecting the features. But using KNN clustering for feature selection makes the authentication schemes complex and it is only applicable for real time applications.

In the next chapter, we propose a new method for authentication of smart phone users based on performing physical activities. A new technique is introduced for selecting the most important features for authenticating users.

CHAPTER III: PROPOSED METHODOLOGIES

This first part of this chapter presents a new strategy for feature selection in order to build an efficient machine learning model for intrusion detection system. The second part of this chapter proposes a smart phone user authentication scheme that authenticates users continuously, based on performing physical activities. The proposed new strategy for feature selection is applied to find the most important features for recognizing users. Different machine learning algorithms are explained and applied for building the continuous user authentication model.

3.1 Research Methodology for Intrusion Detection System (IDS)

The problem of large dimensionality of NSL-KDD requires a feature selection to obtain a better accuracy rate and reasonable model interpretation (Dhanabal1, 2015). There are basic algorithms to reduce the dimensionality of dataset. By using these algorithms the characteristics of the original data is preserved and only nonessential data are removed. According to (Kantardzic, 2011) when basic operations of reducing the datasets are performed, the following parameters can be used to compare what we have lost or gain before feature reduction. The parameters are described below (Kantardzic, 2011):

1) Computing Time: Data reduction is done with the hope of leading to reduction of the time required for the data mining algorithm. However, in some cases the time needed for data reduction is not affordable (Kantardzic, 2011).

2) Predictive/descriptive Accuracy: This is the dominant measure for machine learning models. By removing redundant and irrelevant data, a faster and high accuracy model can be built (Kantardzic, 2011).

3) Representation of the Data Mining Model: Reducing the dimensionality of the data, contributes to building an easier model to be understood, which result in better interpretation. Even if data reduction cause a small tolerable decrease in the accuracy, a balance between the

simplicity of the model and the accuracy is needed. The ideal case is to achieve a reduced time, high accuracy and simplicity representation at the same time with data reduction (Kantardzic, 2011).

In this research, for data reduction, three feature selection techniques, chi-squared, information gain and correlation based, are utilized for a new majority vote system that selects the relevant attributes. In features selection techniques the irrelevant and redundant features will be removed from the data. Feature selection algorithms typically lie in two categories: feature ranking and subset selection (Kantardzic, 2011). Feature ranking scores all features by a specific metric and removes the features that do not achieve a threshold score. While subset selection, searches for optimal subset where features are selected based on ranking (Kantardzic, 2011).

3.1.1 Block Diagram of the Proposed Intrusion Detection System

Figure 1 shows the block diagram of the proposed network intrusion detection. It starts with preprocessing the data and applying different feature selection methods. A number of features are selected and then a voting system is utilized to see which features get the highest votes from all approaches. Figure 2 shows the selecting process for voting system. According to this Venn diagram, the features in region A, get the most votes which means they are the most relevant features for building a model and can be chosen as primary selected features. Any machine learning model can be built based on those selected features. In order to improve the performance, other overlapping regions are investigated. Therefore, the number of features are increased gradually and each time the performance metrics are measured. In other words, a search is done in all overlapping regions and their combinations (A, AB, AC, AD, ABC, ACD, ADB, ABCD) to find the most important features that can contribute to get the highest accuracy. The selected features in the region that gives the highest accuracy, are used for the final machine learning model. A decision tree classifier is used to build a model. Accuracy,

precision, recall and f1-score are used as performance metrics to study the performance of the proposed method. The results are discussed in the next chapter.

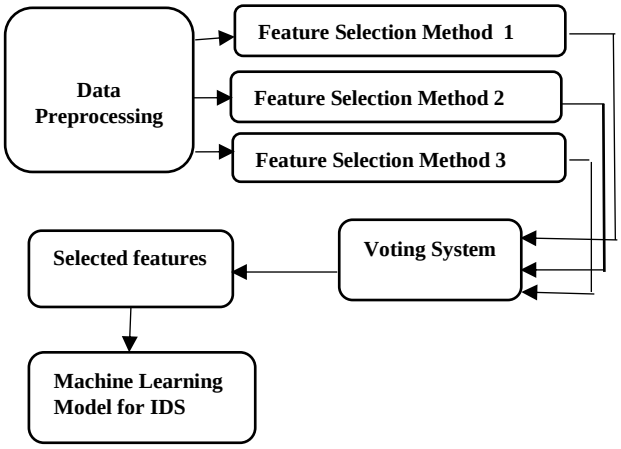


Figure 1 The block diagram of the proposed network intrusion detection system.

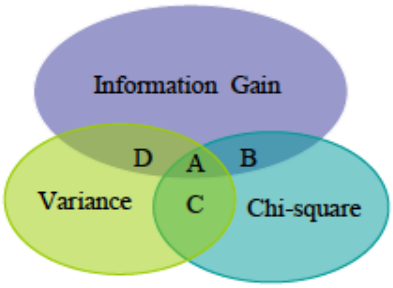


Figure 2 The Venn diagram for voting system

3.1.2 Data Preprocessing

The NSL-KDD dataset includes KDDtrain+.txt and KDDtest+.txt, all the different attack traffic in the dataset is grouped into one class named as an anomaly. KDDtrain consists of 125973 instances. Each record has 41 features. The details of attributes and their descriptions are available in (Tao, 2004). Table 1 summarizes the description of this dataset. There are three types of features, nominal, numeric and binary. Since machine learning methods cannot work on nominal features they are converted to numeric by encoding them using one-hot encoding in Python. The nominal features are "protocol_type", "service", flag". Protocol_type is transferred to 3 new features, service to 70 new features and flag to 11 new features. Therefore the 41-feature data set is transformed to 122 feature dataset. After encoding the KDDtest+

dataset, 116 features is obtained. service_aol, service_aol, service_http_2784, service_http_8001, service_red_i, service_urh_i are the missing values in categorical features that need to be add for testing the classifier without feature selection.

Table 1 Summary of NSL-KDD

Number of instances	**Number of features before categorization**	**Number of features after encoding**	**Type of features**	**Category of features**
125973	41	122	Nominal Binary Numerical	Time related content related host based related

The values of numeric features have different scales and sometimes they are affected by outliers (Ganapathy, 2013) The large valued features may affect the results by some classifier due to having imbalanced values. Therefore, we need to scale the features to give them all equal weight. Normalization is used for scaling with the following formula:

$$x = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

3.1.3 Feature Selection Methods

The three feature selection method we selected for this research are described below.

1) Chi-square: Chi-square test is the measure of dependency between variables. With this function, the most likelihood class-independent and irrelevant attributes for classification are eliminated. The features are ranked by the chi square scores, and the top ranked features for model training are selected. The equation for this test is (Kantardzic, 2011):

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{k}(A_{ij} - E_{ij})^2 / E_{ij} \tag{2}$$

Where, k= number of classes**,** A_{ij} = the number of instances in the ith interval, jth class**,**

E_{ij} = the expected frequency of A_{ij} **,** which is computed as $(R_i C_j) / N$

R_i = the number of instances in the ith interval = $\sum A_{ij}, j=1,...k$

C_j = the number of instances in the jth class= $\sum A_{ij}, i=1,2$

N = The total number of instances = $\sum R_i, i = 1,2$

2) Information Gain: The information gain (IG) evaluates attributes by measuring their information gain with respect to the class (Boujnouni, 2018). The formula is given by:

$$I(c_1,.....c_m) = -\sum_{i=1}^{m} \frac{c_1}{c} \log(\frac{c_1}{c}) \quad (3)$$

Where c_l / c is the probability of a sample belonging to class c_l. And c is the number of data samples with different classes. If a feature F has n different values that divides the training set into v subsets where c_l is the subset corresponds to value f_i for feature F. The entropy of the feature F is:

$$E(F) = \sum_{i=1}^{y} \frac{c_1}{c} \times I(c_i) \quad (4)$$

Information gain for F is defined as:

$$Gain(F) = I(c_1,.....c_m).E(F) \quad (5)$$

3) Variance Threshold: It removes all features whose variance doesn't meet some threshold. It calculates the variance of each feature by then drops the features with variance below the threshold.

$$\sigma^2 = \frac{\sum (x-\mu)^2}{N} \quad (6)$$

Where μ is the mean and N is the number of instances.

3.1.4 Decision Tree Classifier

Decision tree is a structure that consists of leaves, nodes and branches, in which leaves represent classifications and nodes represents a splitting test and the branches are the outcome of the test for splitting the attributes and the links that features lead to those classifications. As a result, to classify an instance, the nodes of the decision tree test its feature values in order to

label them (Buczak, 2016). An example of a simple decision tree with two features X, Y and binary classification is shown in Figure 3 (Kantardzic, 2011).

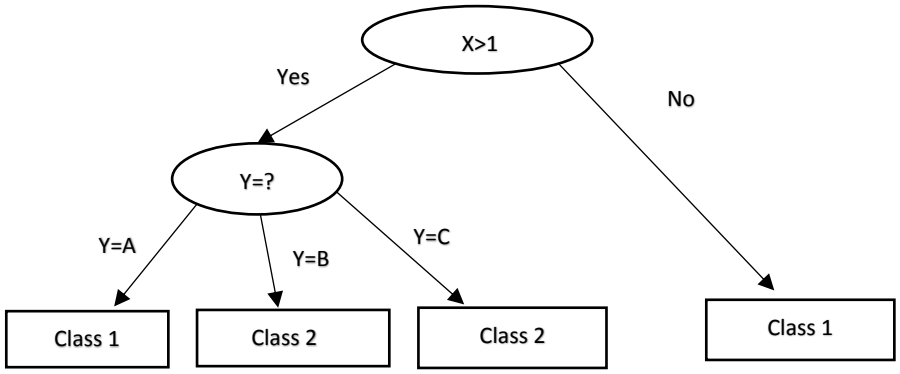


Figure 3 A simple decision tree for a binary classification.

The best-known methods for decision trees are the ID3 and C4.5 algorithms. Both algorithms build decision trees based on information entropy splitting criteria. C4.5 selects the features with the highest gain ratio (difference in entropy) as the splitting criterion and choose the features that splits its set of examples into subsets effectively and then a recursion is done on the smaller subsets until all the training examples are labeled (Buczak, 2016). The formula (5) shows the computation of information gain for splitting criteria.

3.1.4 Performance Evaluation

The above-mentioned methods were applied on NSL-KDD dataset. In order to measure the classification performance, decision tree classifier is used on The KDDtrain for training and KDDtest for testing. To pick the best features for getting the highest accuracy, the proposed voting system is applied. It is important to evaluate the classification process and measure the performance of the algorithm each time a region is investigated. There are different metrics that we used to evaluate the classification algorithms. They are ccuracy, precision, recall, an F-measure that are defined below. Here, TP is true positive, TN is true negative, FP is false positive and FN is false negative they are defined according to Table 2.

Table 2 Confusion matrix foe two-class classification model

Actual Class / Predicted class	Class 1	Class 2
Class 1	True Positive (TP)	False Positive (FP)
Class 2	False Negative (FN)	True Negative (TN)

1) Accuracy: The percentage of predictions that are correct

$$Acuracy = (TP + TN) / (TP + FN + FP + TN) \quad (7)$$

2) Precision: The percentage of correctly classified positive cases to the cases classified as positive:

$$Percision = (TP) \ / \ (TP + FP) \quad (8)$$

3) Recall: The percentage of positive cases that were successfully classified as positive:

$$Recall = (TP) \ / \ (TP + FN) \quad (9)$$

4) F1-Score: Conveys the balance between the precision and the recall. It measures the proportion of positive cases incorrectly classified as negative (Whang, 2018):

$$F1-Score = 2*((precision * recall) / (precision + recall)) \quad (10).$$

3.2 Research Methodology for Authentication of Smart Phone Users

This research proposes an authentication method by utilizing smartphone inertial sensors for recognizing users based on performing activities of daily living including walking, standing, sitting, walking downstairs and upstairs, jogging and biking. The user authentication system includes four main steps: sensing or data collection, preprocessing and feature extraction and training or classification. Figure. 4 shows the block diagram of the proposed system. Each of those steps are explained in details in the following.

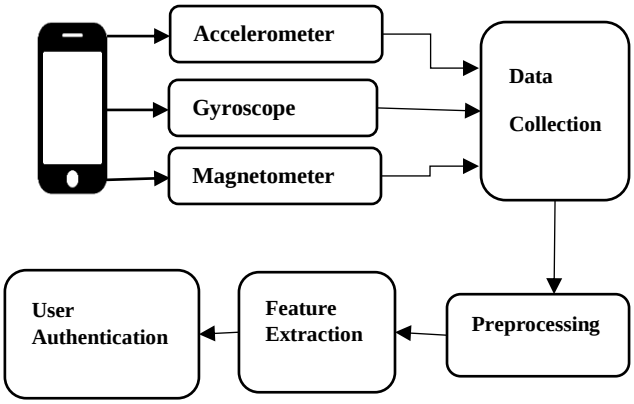


Figure 4 Proposed methodology for smartphone user authentication

3.2.1 Dataset for Smart Phone Users

A public dataset for physical activity recognition is used for this research (Shoaib, 2013). In this dataset, 10 participants performed 7 different daily activities including walking, sitting, standing, jogging, walking upstairs, walking downstairs and biking for three minutes. The participants were male and each of them was equipped with five smart phones at five different position on their bodies including left and right pocket, right wrist which corresponds to holding the smart phone in the right hand, and the waist position which represents a smart phone that is hung on a belt clip (Ehatisham-ul-Haq, 2018). The data were recorded at a rate of 50 Hz from the smartphone inertial sensors including accelerometer, gyroscope, magnetometer to measures acceleration, rotation and magnetic field strength respectively. Each sensor's data is measured along the x-axis, y-axis, z-axis. Previously in (Shoaib, 2014) they showed that accelerometer and gyroscope play the leading role in activity recognition and the combination of them with magnetometer improves the overall performance of activity recognition system. These sensors are sensitive to orientation and this can affects the results of activity recognition algorithms because the sensors reading varies by changing the orientation of smartphones (Ehatisham-ul-Haq, 2017). To address this issue, magnitude of the sensors are added as an orientation independent feature according to equation 11.

$$Magnitude = \sqrt{x^2 + y^2 + z^2} \tag{11}$$

As a result, each sensor's data has four dimensions (x, y, z, Mag)

3.2.2 Preprocessing

The collected data needs to be processed for two main reasons: to remove the noise and to segment the data for feature extraction.

3.2.2.1 Noise Removal

Noise can damage the useful information in sensors inertial signal. In order to remove the noise, an average smoothing filter that is applied in (Su, 2014) is adopted. This filter takes the average of the two adjacent data to eliminate the sudden spike that might happen if the user drops the smartphone.

3.2.2.2 Data Segmentation

Another important preprocessing step is to divide the signal data into small segments for feature extraction and training the machine learning models. There are two categories of segmentation: overlapping segmentation and no-over lapping segmentation. The fixed size no-overlapping window segmentation is the most common method in activity recognition systems since it makes the segmentation less computational and is capable of retrieving data continuously over time (Su, 2014). According to (Su, 2014), the size of the window is very important on the final accuracy of recognition. Previous studies on activity recognitions showed that a window size of a time interval of 5 second is enough to recognize the activities (Shoaib, 2013), (Anjum, 2013). As a result, a fixed-size window of 5 second with no overlapping between the samples was selected for segmenting the data for every sensor along each axis.

3.2.3. Feature Extraction

In preprocessing phase, various features are extracted from the raw sensor data for training and testing of classification method. There are two basic types of features, time domain and frequency domain. The time domain features are used more common in activity recognition.

The main reason is that frequency domain features are computationally complex due to Fast Fourier Transformation (FFT) (Shoaib, 2013), (Anjum, 2013). The selection of features is also an important factor in activity recognition. According to (Shoaib, 2015), the number and type of feature is a design decision. For this reason, it is important to analyze the addition of a feature in improvement of the performance of the activity recognition system (Shoaib, 2015). In many studies some features are added without evaluating their impact (Shoaib, 2015). One of the most common solutions is to begin with a simple set of features and add the new features and examine how they improve the performance. In a research (Shoaib, 2014), four feature sets that consists of at most four features are selected. However, the number of features they investigated are small sets of features. In another study (Quiroz, 2017), they conducted several experiments on dataset that include 561 features extracted from a human activity recognition public dataset (HAR). They compared various feature sets and analyzed how those sets influence the accuracy of different classifiers to find the best feature sets. However, this method requires a series of experiments for selecting different feature sets and applying classifiers for all of those selected sets and making a decision on the final feature sets. Another way of feature selection is to uses multilevel features in which the data is first clustered (Bulling, 2014). An example of this method is in (Ehatisham-ul-Haq, 2017) where k-mean clustering was used for feature selection on a window segment. However this method makes the feature selection more computational by using KNN clustering. In our study to find effective, yet smaller feature sets, the novel voting system that was introduced in this chapter for feature selection is applied. With this method the best features are selected by letting the different scientific feature selection techniques make the final decision and determine which of them are more important in user authentications. In order to implement this method, some features that have been used in recent studies on activity recognition are used for the voting system. All of these features are extracted over a fixed size window of 5 second. The features and their definitions are described below.

The most common used features in time domain are mean and variance/standard deviation of the sensor data. They are widely used in activity recognition using sensors in smart phones along with other time domain or frequency domain features (Shoaib, 2013), (SU, 2014), (Anjum, 2013), (Sun, 2010), (Anguita, 2013). They are defined as:

Mean: It is the average of sample values over a window of data samples

$$\mu = \frac{1}{T}\sum S(t) \quad (12)$$

Where T is the window segment size.

Variance/standard deviation: Variance (σ^2) is the average of the squared differences from the mean. The standard deviation is the square-root of the variance σ.

$$\sigma^2 = \frac{1}{T}\sum (S(t) - \mu) \quad (13)$$

Median: The median is the separator of the higher half of the data from the lower half (Ehatisham-ul-Haq, 2018), (Ehatisham-ul-Haq, 2017), (Shoaib, 2014), (Figo, 2010).

Maximum amplitude: It is the maximum value over a window segment in each dimension. (Ehatisham-ul-Haq, 2018), (Ehatisham-ul-Haq, 2017), (Shoaib, 2014), (Figo, 2010),(Anguita, 2013)

$$S_{max} = max\big(S(t)\big) \quad (14)$$

Minimum amplitude: It is the minimum value over a window segment in each dimension. (Ehatisham-ul-Haq, 2018), (Ehatisham-ul-Haq, 2017), (Shoaib, 2014), (Figo, 2010), (Anguita, 2013)

$$S_{min} = min\big(S(t)\big) \quad (15)$$

Range (peak to peak signal value): It is defined as the difference between maximum and minimum of a signal (Ehatisham-ul-Haq, 2017), (Anjum, 2013)

$$S_{pp} = S_{max} - S_{min} \quad (16)$$

Root Mean Square (RMS): For a signal s_i that represents n discrete values $\{s_1, s_2, \ldots, s_n\}$, RMS is obtained using equation (17): (Shoaib, 2014), (Figo, 2010)

$$RMS = \sqrt{\frac{{s_1}^2 + {s_2}^2 + \ldots + {s_n}^2}{n}} \quad (17)$$

Kurtosis: If m_2 and m_4 are the 2nd and 4th moment from the mean then: (Ehatisham-ul-Haq, 2018), (Ehatisham-ul-Haq, 2017), (Shoaib, 2015)

$$K = \frac{m_4}{{m^2}_2} \quad (18)$$

Skewness: If m_3 is the 3rd moment about the man then (Boujnouni, 2018), (Ehatisham-ul-Haq, 2018), (Shoaib, 2015):

$$S = {m_3}\Big/{m_3^{\frac{2}{3}}} \quad (19)$$

Peak to peak time: The time that is needed to go from the minimum values to the maximum value of a signal over a window segment (Ehatisham-ul-Haq, 2018), (Ehatisham-ul-Haq, 2017):

$$t_{pp} = t_{s_{max}} - t_{s_{min}} \; where \; t_{s_{max}} = t \; l \; s(t) = s_{max} \; and \; t_{s_{min}} = t \; l \; s_{min} \quad (20)$$

Peak to peak slop: The ratio of maximum amplitude to the peak to peak time (Ehatisham-ul-Haq, 2018), (Ehatisham-ul-Haq, 2017).

$$s_{pps} = \frac{s_{pp}}{t_{pp}} \quad (21)$$

Absolute latency to amplitude ratio (ALAR): Absolute latency to amplitude ratio (Ehatisham-ul-Haq, 2018), (Ehatisham-ul-Haq, 2017)

$$ALAR = \left|\frac{t_{s_{max}}}{s_{max}}\right| \quad (22)$$

Signal correlation: To calculate the correlation for sensor signals, it is necessary to calculate correlation between each pair of axes of the sensor data (Su, 2014), (Figo, 2010), (Feng, 2015). The most common used is the Pearson's product -moment coefficient according to the following formula (Figo, 2010):

$$\rho_{x,y} = \frac{cov(x,y)}{\sigma_x \sigma_y} \quad (23)$$

Zero-crossing: The number of point where a signal passes a specific value that is half of the signal range (Shoaib, 2014), (Figo, 2010). In our study it is the mean of the window segment is considered for that value (Shoaib, 2014).

Spectral Energy: The spectral energy of a signal can be computed as the square sum of its discreet FFT (Fast Fourier Transform) coefficient normalized by length the sample window (Ehatisham-ul-Haq, 2017), (Shoaib, 2014), (Su, 2014), (Anguita, 2013), (Figo, 2010), (Sun, 2010):

$$E(f) = \sum |S(f)|^2 /_T \quad (24)$$

Where $S(f)$ is the discrete Fourie transform.

Entropy: Entropy is computed by the normalize information entropy coefficient magnitudes excluded DC component (Figo, 2010). The DC component is the first coefficient in the spectral of a signal and it is much larger than the other spectral coefficients (Figo, 2010). The equation shows the formula for entropy (Ehatisham-ul-Haq, 2017), (Shoaib, 2015), (Anjum, 2013):

$$H\big(S(f)\big) - \sum P_i\big(S(f)\big) \log_2 \Big(P\big(S(f)\big)\Big) \quad (25)$$

Where P is:

$$P(f) \frac{E(f)}{\sum_i E(i)} \quad (26)$$

Sum of FFT coefficient: This is defined as the summation of the some number of FFT coefficients (Figo, 2010). The first five FFT coefficients are selected in our study (Shoaib, 2014).

3.2.4. User Authentication:

After feature extraction, the next step is to propose a user authentication method to identify a smartphone user as authenticated or not authenticated. Hence a suitable classifier needs to be chosen to user authentication schemes. The first experiment is to recognize the ten different

participants doing seven different activities. For this purpose, four different classifiers are select for training the data set. SVM, Decision tree, KNN, Random forest are used. The decision tree classifier is described in section 3.4. Random forest, SVM and KNN classifiers are explained in the following.

3.2.4.1 Random Forest Classifier

Random forests are multi-class classifiers with a fast and high effective performance. It is an ensemble of n number of trees which include split and nod leaves. Each tree is trained on randomly selected of a data set. The output of this classifier is the mode of that is the mode of the classes of the individual trees or mean prediction (regression) of the individual trees. Figure 5. displays a random forest classifier.

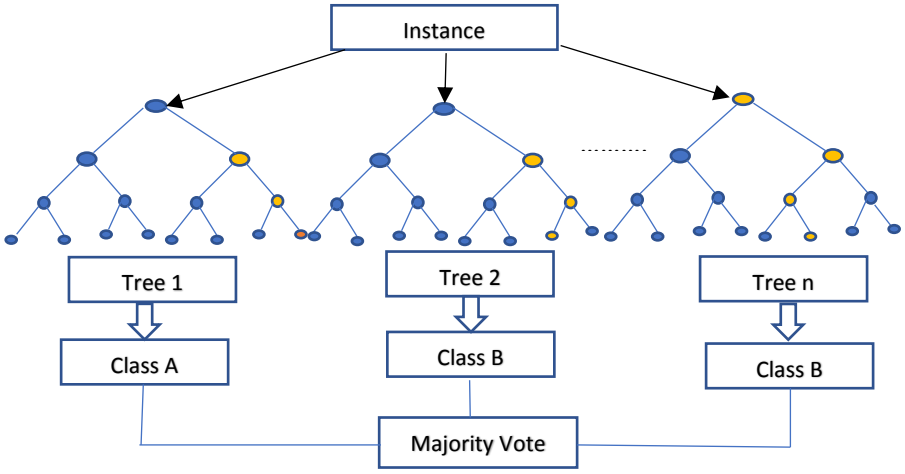


Figure 5 Random forest classifier for binary classification

3.2.4.2 Support Vector Machines (SVM) Classifier

According to (Buczak, 2016), "the SVM is a classifier based on finding a separating hyperplane in the feature space between classes in such a way that the distance between the hyperplane and the closest data points of each class is maximized. The approach is based on a minimized classification risk rather than on optimal classification. SVMs are well known for their generalization ability are particularly useful when the number of features, m, is high and the number of data points, n, is low (m >> n). Various types of dividing classification surfaces can be realized by applying a kernel, such as linear, polynomial, Gaussian Radial Basis

Function (RBF), or hyperbolic tangent. SVMs are binary classifiers and multi-class classification is realized by developing an SVM for each pair of classes." In Support Vector Machine, we have a set of observations and we want to classify them or find out which class they belong to. So a boundary that separate between the classes needs to be found. The boundary line is searched through the maximum margin. The main object of SVM is to find the best decision boundary line that will help us separate our classes. Kernel function is used for this purpose. In general kernel is a function of similarity (it measures the similarity between two data points). It has two inputs and spits out how similar they are. Figure 6 shows a SVM classifier for a binary classification for two dimensional dataset for a linear kernel.

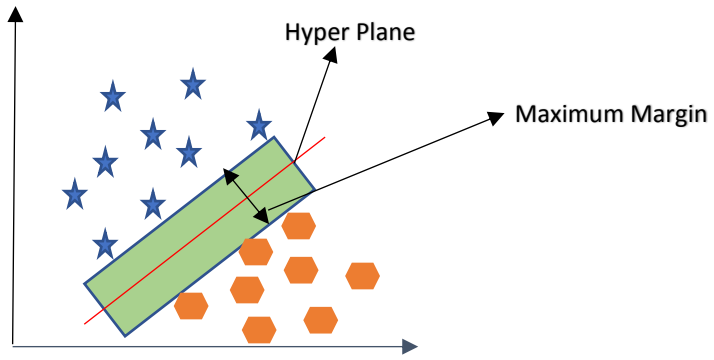


Figure 6 SVM classifier for binary classification

An example of a kernel function (Gaussian Radial Basis Function (RBF) is as follows

$$k\left(x, \vec{l_i}\right) = e^{\frac{-\|x-\vec{l_i}\|}{2\delta^2}} \quad (27)$$

Where k stands for kernel, x vector is some points in the data set, l is landmark and the i means there may be several landmarks, $\| x - \vec{l_i} \|$: means the difference between x and l.

δ : is a fixed parameter that we decide on

3.2.4.3 K Nearest Neighbors (KNN)

K-Nearest neighbor is one of the most commonly used algorithms for pattern recognition. The algorithm gets a feature vector from the input data and assigns it to its nearest neighbor which can be a class protype or a feature vector from the training set. The nearest neighbor is

determined by calculating the distance between the feature vectors. Different distance measures such as Manhattan, Minkowski and Euclidean distance is used but Euclidean is usually the default one. Number of k neighbors (for example: k=5) can be specified. The k nearest neighbors of the new data point according to distance measures are calculated. Among these k neighbors, the number of data points in each category (class) are counted and the new data point is assigned to the category that has the most neighbors. Figure 7 displays an example of KNN classifier.

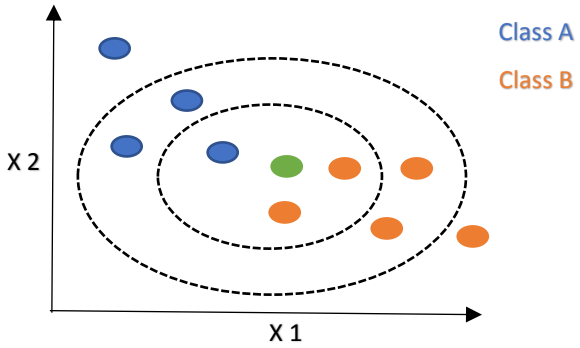


Figure7 KNN classifier

The next chapter discusses the results of the proposed methodologies for network intrusion detection and authentication of smart phone users.

CHAPTER V: EXPERIMENTAL RESULTS

The first part of this chapter discusses the experimental results of the proposed methodology for intrusion detection system. The second part of this chapter discusses the experimental results of the proposed methodology for authentication of smart phone users based on performing physical daily activity.

5.1 Experimental Results for the Proposed Intrusion Detection System

The proposed method is applied for intrusion detection system for NSL-KDD dataset. And the results are reported. Without feature selection the accuracy of the decision tree model is 79.96%. The feature sets resulting from information gain gives an accuracy of 79.91% with having $IG \geq 0.02$. The subsets consist of 31 features out of 122. Chi square method has the accuracy of 79.91% with 20 features. The best threshold for variance is set at 0.01 using trial and error method that gives the accuracy of 75.30% with 18 features. The accuracy, precision, recall and f1-score and the number of features (no. feature) are reported in Table 3. It is shown that the three different methods select 16 common features. The decision tree classifier was built initially based on those selected features in region A which is the intersection of the three selected sets from the three feature selection methods. The output of this classifier gives an accuracy of 76.76%. To increase the accuracy, the number of features are increased by searching through other overlapping regions (A&B, A&C, A&D) and the accuracies and the number of features are recorded in Table 3. The results show that although the accuracy in A&C,A&D are not increased in comparison with without feature selection, but the number of attributes have dropped significantly which make the model much simpler. In A&B region, not only the number of features are reduced but also the accuracy have raised. The search can be continued for other regions (A&B&C, A&C&D, A&D&B, A&B&C&D) to find if the accuracy can be raised with less number of features compare to a model without feature selection. According to this table, by using decision tree the best results come from the intersection of

information gain and chi-square scores with a significantly less number of features. It is also shown that the accuracy increases in most overlapping regions with a few number of features compared to models that are built with from only one method of feature selection. Therefore, building the model based on those features will lead to a more efficient model. As a result, the those features in those regions are the most relevant ones for making a machine learning model for an IDS. To visualize the accuracy of different regions, a bar chart is displayed in Figure 8. All The 20-selected feature for each of the methods are mentioned in Table 4. The details about those feature are discussed in (Dhanabal, 2015).

Table 3 Evaluation of the proposed method

Regions	**Accuracy**	**Precision**	**Recall**	**F1-score**	**no. features**
A	76.76%	83%	77%	77%	16
A&B	80.6%	96%	69%	80%	20
A&C	76.74%	83%	77%	76%	20
A&D	76.63%	83%	77%	76%	21
A&B&C	80.59%	96%	69%	80%	20
A&C&D	76.61%	95%	62%	75%	17
A&D&B	75.73%	91%	64%	76%	21
A&B&C&D	76.24%	80%	76%	76%	21
Information Gain	79.9%	83%	80%	80%	31
Chi-square	79.91%	85%	80%	80%	20
Variance	75.3%	82%	75%	75%	18

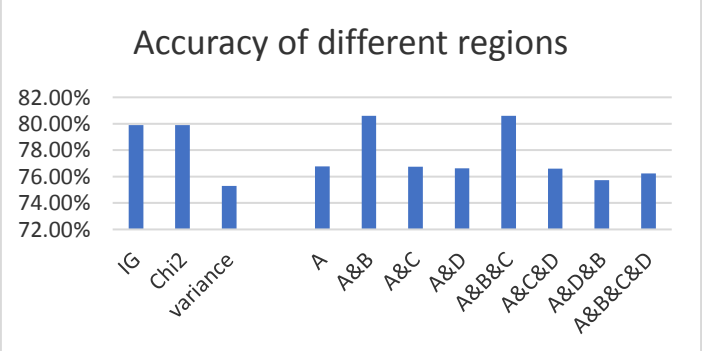


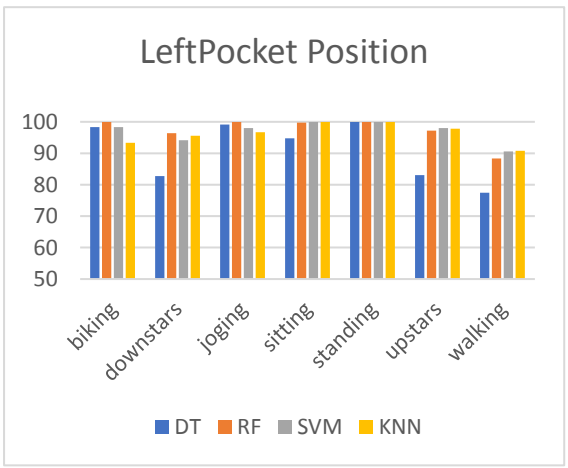
Figure 8 The accuracy of different regions

Table 4 Selected features

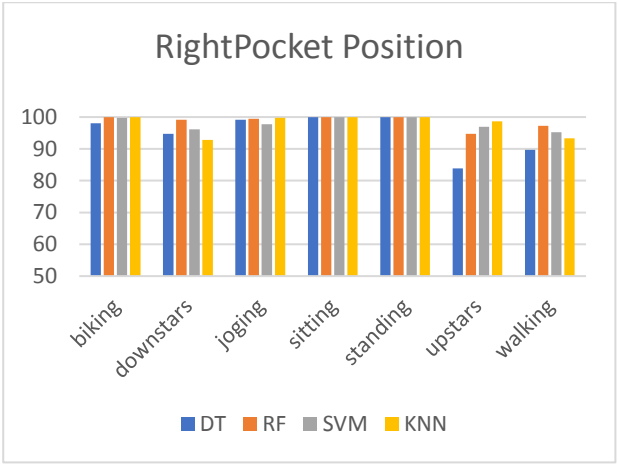
Selected features:	
flag_S0	same_srv_rate
dst_host_rerror_rate	service_private
service_http	count
serror_rate	dst_host_srv_serror_rate
dst_host_srv_rerror_rate	dst_host_srv_count
dst_host_serror_rate	flag_SF
rerror_rate	protocol_type_udp
service_domain_u	srv_rerror_rate
logged_in	srv_serror_rate
dst_host_same_srv_rate	service_smtp

5.3. Experimental Results for Activity Recognition

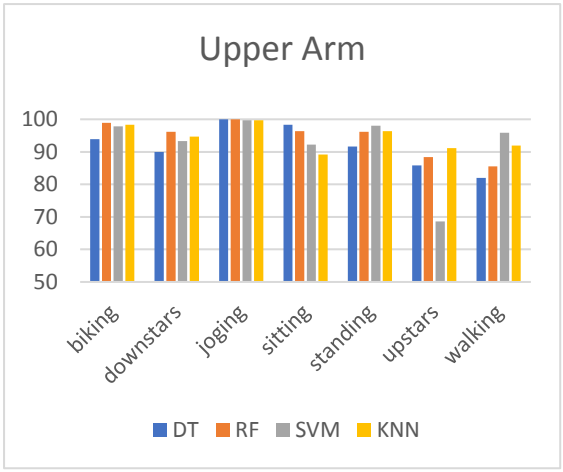
Several experimental results were conducted to study the performance of the new voting system for feature selection. All of features in section 3.2.3 are extracted from each axis of these sensor signals over the sample window segment. The voting system is applied on the features to find which features get the most votes for having the highest accuracy for the model. First we applied the proposed feature selection strategy for activity recognition and then we apply it on user authentication. The experimental results for results activity recognition are reported below.



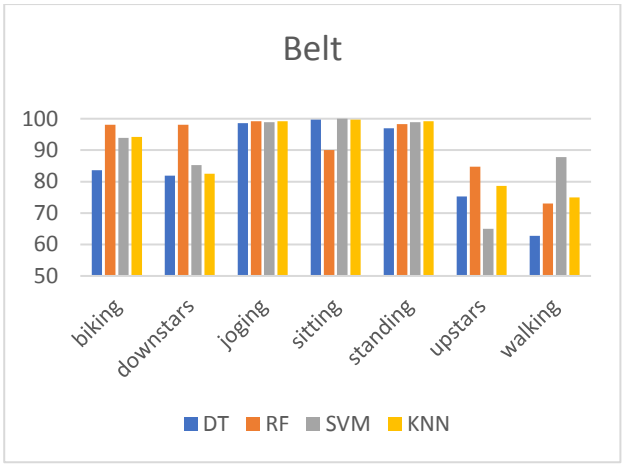
(a)



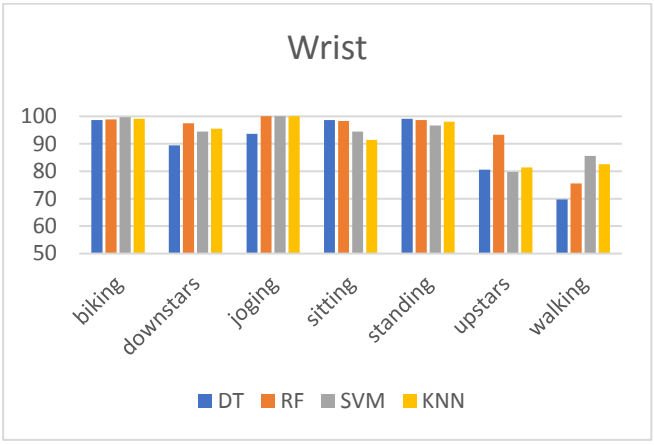
(b)



(c)



(d)



(e)

Figure 9 Individual accuracies for four different classifiers in five different body positions

Figure 9 shows that the individual accuracies for standing and sitting and jogging are higher than other activities in all body position. The accuracy of waking depends on different body position and this activity is less recognizable than the other ones. SVM is the best classifier for recognizing walking in almost all body position. For the other activities, overall random forest gives the highest accuracies in all body positions.

With the voting system, we can also analyze the features that are selected by this system for classifiers. This technique for feature selection allows us to analyze each sensor. As an example, Figure 10 shows the features that are used for random forest classifiers for left pocket position. We can see that for variance, peak to peak signal value, peak to peak slope and

skewness features are used for all sensors for four dimensions (12 features). For correlation only one feature was used, and it is the correlation between y axis and z axis for accelerometer (corr_acc_y_z). Kurtosis is only use for gyroscope in x axis. Alar was also use for accelerometer only in z axis (kurt_acc_x_z). For all other features we can also investigate for which axis they are important to be computed.

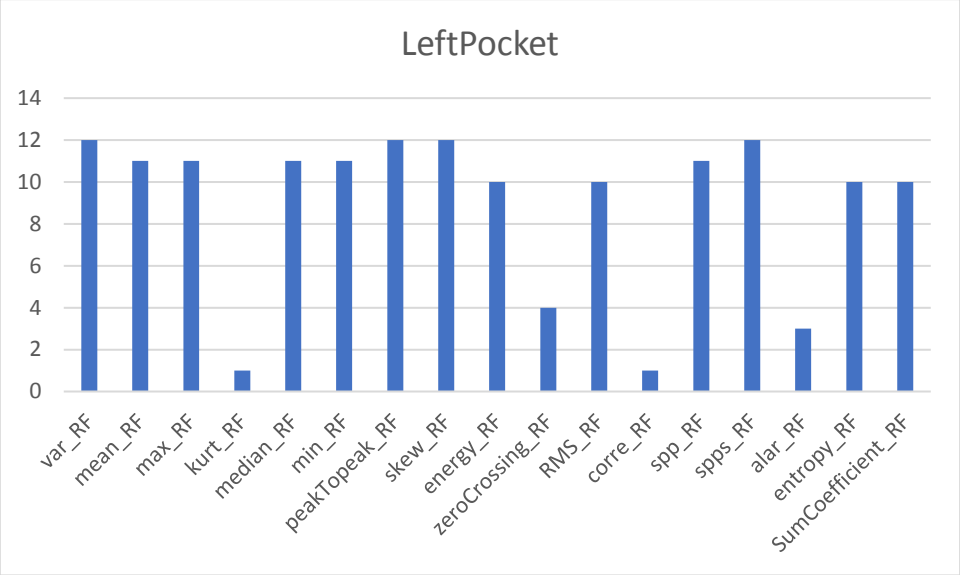


Figure10 The features that are selected by voting system for left pocket position

5.4 Experimental Results for the Proposed Smart Phone Authentication

The dataset is labeled for seven activities and the users who perform them are also available. Therefore, we have a ground table for the activities and also users. Our idea is to recognize the users directly, from this labeled activity data. For this reason, A 10-fold cross validation method is used for evaluation of the model. According to this method, the data set is split randomly into ten sets and iterates 10 times so that every set is used for training and testing the classifiers. The results are the average of these 10 repetitions. It generally results in a less biased or less optimistic estimate of the model skill than other methods, such as a simple train/test split. Table V shows the average performance for all ten participants for five body position with decision tree (DT), random forest (RF), support vector machine (SVM), and K nearest neighbor (KNN) classifiers. The performance metric for evaluating the classifiers is accuracy, precision, recall

and F-measure. These metrics are explained in detail chapter 3.1.4. An entropy index is used for splitting the nods for DT classifier. A RBF kernel is used for SVM classifier and multiclass classification is handled according to a one-vs-one scheme. For KNN classifier K=7 is set and Euclidean distance metric is used. For random forest the number of decision trees are set to 50. Table 5 shows the performance parameters of the selected classifiers for user authentication at five different body position. It is observed that random forest classifiers provide the best accuracy compared to other classifiers for all body position. Moreover, other metrics are also high with this classifier in all body position. However, SVM classifiers provide lowest accuracy for all metrics in all body position. KNN classifier performance is the second-best classifiers for all position. Its accuracy is very close to random forest. However, this classification is not practical one the number of data increases. According to the Table 5, the right pocket and left pocket positions gives the best accuracy scores. Therefore, recognizing users is easier if the phone is in their right or left pocket. The results state that having the phone in belt position make the authentication of users more difficult than the other positions. The number of features that are selected from voting system and are used in classification are also reported, these features can be recorder and be applied for recognizing new individuals by collecting raw data from the motion sensors in smartphones. In real time, if a person performs an activity that is unknown for system, the proposed system can be used for training the new collected data from the sensors and extract the important features from the recorded features and as a result adjust its self to identify users.

Typically, the owner of a cell phone is one person who has full access to everything in the smart phone. An owner may share his/her phone with other people, for performing any of their tasks the phone owner allows them (Ehatishum-ul-Haq, 2018), (Sitova, 2016). These people are called supplementary. Hence, there are three labels for user authentication.

Table 5: Performance measures of classifiers for 10 different user authentication

Left Pocket					
Classifier	Accuracy	Precision	Recall	f1-score	No. F
DT	83.52%	84.14%	83.52%	83.44%	97
RF	92.23%	92.73%	92.23%	92.18%	97
SVM	77.51%	79.55%	77.55%	77.15%	41
KNN	84.13%	86.11%	84.13%	83.73%	65
Right Pocket					
Classifier	Accuracy	Precision	Recall	f1-score	No. F
DT	83.78%	84.91%	83.78%	83.6%	92
RF	92.05%	92.38%	92.05%	91.97%	78
SVM	72.13%	74.49%	72.13%	71.44%	92
KNN	82.34%	84.89%	82.34%	81.83%	113
Wrist					
Classifier	Accuracy	Precision	Recall	f1-score	No. F
DT	85.54%	85.57%	85.54%	85.28%	51
RF	92.61%	93.14%	92.61%	92.57%	77
SVM	77.51%	78.85%	77.55%	76.35%	77
KNN	84.34%	86.03%	84.34%	84.06	77
Upper Arm					
Classifier	Accuracy	Precision	Recall	f1-score	No. F
DT	86.16%	86.90%	86.16%	85.97%	96
RF	92.08%	93.32%	92.08%	91.92%	96
SVM	74.97%	76.12%	74.97%	74.46%	56
KNN	87.51%	88.70%	87.51%	87.39%	56
Belt					
Classifier	Accuracy	Precision	Recall	f1-score	No. F
DT	84.30%	85.50%	84.30%	84.11%	120
RF	89.23%	89.96%	89.23%	89.16%	103
SVM	78.74%	80.97%	78.74%	78.38%	66
KNN	89.30%	90.02%	89.30%	89.23%	111

Unauthenticated, supplementary and authenticated. For this three labeled classification the user authentication is applied, and the results are reported in the table below. To get the best individual accuracy for each user class, a balanced data set is needed. To make this data set balanced, one participant was selected randomly as a supplementary and another one as authenticated. The other users are considered as unauthenticated and a certain number of records are selected for each of them to have a balanced dataset for training. According to Table 6, the best accuracy is from random forest as well. The performance of SVM has the least performance. Moreover, the individual accuracy for this classifier are not acceptable. This table also approves that the highest accuracy is from left pocket and the worst come from the upper arm.

Table 6 Performance measures and individual accuracies of different classifiers for user three labeled authentication

Left Pocket					
Classifier	Accuracy	Precision	Recall	f1-score	No. F
DT	83.515%	84.756%	83.528%	83.129%	116
RF	93.378%	93.798%	93.384%	93.358%	85
SVM	78.517%	79.098%	78.533%	78.161%	110
KNN	87.304%	88.673%	86.427%	85.942%	109
Individual accuracy of Left Pocket					
Classifier	authenticated		supplementary	unauthenticated	
DT	82.53%		92.06%	83.03%	
RF	90.87%		98.80%	90.62%	
SVM	69.19%		98.41%	67.57%	
KNN	95.23%		99.6%	71.73%	
Right Pocket					
Classifier	Accuracy	Precision	Recall	f1-score	No. F
DT	81.268%	83.803%	81.348%	80.881%	106
RF	90.335%	92.796%	90.384%	90.015%	138
SVM	60.391%	66.826%	60.492%	57.684%	117
KNN	80.029%	84.051%	80.199%	78.628%	98
Individual Accuracy of Right Pocket Position					
Classifier	authenticated		supplementary	unauthenticated	
DT	86.50%		80.15%	77.34%	
RF	90.47%		90.85%	90.73%	
SVM	66.66%		67.02%	47.65%	
KNN	94.85%		92.28%	53.5%	
Wrist					
Classifier	Accuracy	Precision	Recall	F1-score	No. F
DT	79.668%	80.646%	78.612%	78.949%	97
RF	88.232%	89.399%	88.312%	88.136%	68
SVM	62.910%	69.110%	63.0 %	62.027%	100
KNN	80.146%	81.632%	80.312%	79.494%	63
Individual accuracy of Wrist Position					
Classifier	authenticated		supplementary	unauthenticated	
DT	80.95%		72.22%	82.86%	
RF	93.65%		90.87%	80.46%	
SVM	74.60%		68.25%	46.09%	
KNN	87.69%		91.66%	61.32%	
Upper Arm					
Classifier	Accuracy	Precision	Recall	F1-score	No. F
DT	81.826%	82.624%	81.851%	81.391%	108
RF	91.079	91.905%	91.076%	91.054%	91
SVM	48.780%	46.121%	48.738%	45.109%	96
KNN	78.016%	81.6523%	77.244%	75.698%	94
Individual accuracy of Upper Arm Position					
Classifier	authenticated		supplementary	unauthenticated	
DT	84.52%		78.79%	82.78%	
RF	90.04%		92.46%	90.23%	
SVM	78.3%		15.93%	51.95%	
KNN	86.9%		88.49%	56.52%	
Belt					
Classifier	Accuracy	Precision	Recall	F1-score	No. F
DT	90.204%	91.157%	90.241%	90.227%	120
RF	94.339%	95.276%	94.381%	94.062%	116
SVM	74.136%	76.195%	74.158%	71.903%	84
KNN	90.254%	90.254%	90.659%	89.773%	120
Individual accuracy of Belt Position					
Classifier	authenticated		supplementary	unauthenticated	
DT	84.12%		97.61%	89.43%	
RF	88.49%		100%	94.64%	
SVM	79.76%		74.60%	68.35%	
KNN	93.25%		99.2%	76.78%	

VI: CONCLUSION AND FUTURE WORK

5.1 Conclusion

This thesis proposed novel techniques for network intrusion detection system (IDS) and authentication of smartphone users.

For the problem of intrusion detection, we focused on feature selection part of an intrusion detection system (IDS). A novel method is proposed to reduce the complexity of the IDS by reducing the number of features significantly and improve the performance of decision tree classifier. The initial results on NSL-KDD dataset is promising and illustrate that feature subset identified by the overlapping region of information gain and chi-square selects the best features for building an efficient machine learning model for IDS.

This thesis proposed a novel method to authenticate smartphone users directly based on performing daily activity using built-in sensors. Seven activities of daily life including walking, running sitting standing walking upstairs and walking downstairs and biking are used to validate different users. A novel feature section technique is applied to find the most important features in recognizing users for building a machine learning model. For each person, five different positions are employed for keeping a smartphone on the body. It is shown that the performance of user authentication depends on the position of smartphone on the body. A user can easily be recognizing if he/she put the smartphone in the right and left pocket. Four different machine learning algorithms i.e. decision tree, random forest k-nearest neighbors and support vector machine are used for the purpose of user authentication. It is observed that random forest classifier provides the best performance for user authentication. As a result, it is an ideal choice.

5.2 Future Work

Future work for intrusion detection will include developing and applying other feature selection approaches such as principle component analysis (PCA) and other machine learning

classifiers to improve the results of the model. This method can also be utilized for other security datasets such as authentication.

For future work for smart phone user authentication will also include applying PCA method for feature selection. An un supervised machine learning approach can be introduced for adaptive user authentication as the behavior of the user may vary in different ways.

REFERENCES

Aggarwal, P., Sharma S. K. (2015). Analysis of KDD dataset attributes - class wise for intrusion. Procedia Computer Science, 57, 842- 851.

Alzubaidi, A., Kalita, J. (2016). Authentication of smartphone users using behavioral biometrics. IEEE 499 Commun. Surv. Tutorials, 18, 1998–2026.

Anguita, D., Ghio, A., Oneto, L., Parra, X., & Reyes-Ortiz, J.L. (2013). A public domain dataset for human activity recognition using smartphones. In Proceedings of the 21th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Bruges, Belgium, 24–26.

Anjum, A., Ilyas, M.U. (2013). Activity recognition using smart phone sensors. In Proceedings of the 2013 IEEE 10th Consumer Communications and Networking Conference, Las Vegas, NV, USA, 914–919.

Assi1, J.H, Sadiq A.T. (2017). NSL-KDD dataset Classification using five classification methods and three feature selection strategies. Journal of Advanced Computer Science and Technology Research, 7(1), 15-28.

Bulling, A., Blanke, U., & Schiele, B. (2014). A tutorial on human activity recognition using body-worn inertial sensors. ACM Comput. Surv. 1, 1–33.

Boujnouni, M., Jedra, M. (2018). New intrusion detection system based on support vector domain description with information gain metrics. International Journal of Network Security, 20(1), 25-34.

Buczak, A.L, Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE communications surveys & tutorials. 18(2).

Centeno, M.P, Van Moorsel, A., & Castruccio, S. (2017). Smartphone continuous authentication using deep learning autoencoders. In Proc. Int. Conf. Privacy, Secur. Trust (PST), Calgary, AB, Canada 1-9.

Chae, H., Jo, B., Choi, S., Park, T. (2013). Feature selection for intrusion detection using NSL-KDD. Recent Advances in Computer Science. 184–187.

Dhanabal1, L., Shantharajah, S.P. (2015). A study on NSL-KDD dataset for intrusion detection system based on classification algorithms. International Journal of Advanced Research in Computer and Communication Engineering, 4(6).

Ehatisham-ul-Haq, M., Awais Azam, M., Loo, J., K. Shuang, S. Islam, U. Naeem & Y. Amin. (2017). Authentication of smartphone users based on activity recognition and mobile sensing. Journal of Sensors, 17, 2043.

Ehatisham-ul-Haq, M., Awais Azam, M., Naeem, U., Amin, Y., & Loo, L. (2018). Continuous authentication of smart phone users based on activity pattern recognition using passive mobile sensing. Journal of Network and Computer Applications.109, 24-35

Ehatisham-ul-Haq, M., Awais Azam, M., Naeem, U., Amin, Y., Ur Rehman, S., & Khalid, A. (2017). Identifying smartphone users based on their activity patterns via mobile sensing. The 8th International Conference on Emerging ubiquitous Systems and Pervasive Networks, 13, 202-209.

Feng, Z., Mo, L., Li, M. (2015). A random Forest-Based ensemble method for activity recognition. In Proceedings of The 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Milan, Italy, 5074–5077.

Figo, D., Diniz, P.C, Ferreira, D.R. (2010). Preprocessing techniques for context recognition from accelerometer data. Pers Ubiquit Comput, 14, 645-662.

Ganapathy, S., Kulothungan, K., Muthurajkumar, S., Vijayalakshmi, M., Yogesh, P., Kannan, A., (2013). Intelligent feature selection and classification techniques for intrusion detection in networks: a survey. Journal on Wireless Communications and Networking. 2013:271.

Kantardzic, M. (2011). Data mining, concepts, models, methods and algorithms. University of Louisville, 2th edition.

Kumar, K., Batth, J.S. (2016). Network intrusion detection with feature selection techniques using machine learning Algorithms. International Journal of Computer Application, 150, 12.

Kunz, M., Kasper, K., Reininger, H., Möbius, M., & Ohms, J. (2011). Continuous speaker verification in real time. In Proceedings of the Special Interest Group on Biometrics and Electronic Signatures, Darmstadt, Germany, 79–88.

lakhina1, S., S. Joseph & B. verma. (2010). Feature reduction using principal component analysis for effective anomaly–based intrusion detection on NSL-KDD. International Journal of Engineering Science and Technology. 2(6), 1790-1799.

Li, Z., Li, Y., Xu, L. (2011). Anomaly intrusion detection method based on k-means clustering algorithm with particle swarm optimization. In Proceeding of the International Conference on Information Technology, Computer Engineering and Management Sciences, 157-161.

Meng, W., Li, W., Wong, D.S, & Zhou, J. (2016). TMGuard: A touch movement-based security mechanism for screen unlock patterns on smartphones: use this paper for introduction. The 14th International Conference on Applied Cryptography and Network Security (ACNS).

Muaaz, M., Mayrhofer, R. An analysis of different approaches to gait recognition using cell phone based accelerometers. In Proceedings of International Conference on Advances in Mobile Computing & Multimedia. ACM, 293, 2013.

Mukherjeea, S., Sharmaa, N. (2012). Intrusion detection using Naive Bayes classifier with feature reduction. Procedia Technology. 4, 119 – 128.

Neverova, N., Wolf, C., Lacey, G., Fridman, L., Chandra, D., Barbello, B., & Taylor, G. (2016). Learning human identity from motion patterns. IEEE Access, 4, 1810–1820.

Nguyen, H., Franke, K., & Petrovic, S. (2010). Improving effectiveness of intrusion detection by Correlation feature selection. International Conference on Availability, Reliability and Security, IEEE, 17-24.

Panda, M., Abraham, A., & Patra, M.R. (2010). Discriminative multinomial Naïve Bayes for network intrusion detection. Sixth International Conference on Information Assurance and Security.

Pandaa, M., Abrahamb, A., & Patra, M.R. (2012). A hybrid intelligent approach for network intrusion detection. Preceding Engineering, 30, 1-9.

Panda, M., Patra. M.R. (2009). Mining association rules for constructing a network intrusion detection model. International journal of applied engineering research, 4(3), 381-98.

Papamartzivanos, D., Damopoulos, D., & Kambourakis, G. (2014). A cloud-based architecture to crowdsource mobile app privacy leaks. In Proceedings of the 18th Panhellenic Conference on Informatics. ACM, 1–6.

Quiroz, J.C., Banerjee, A., Dascalu, S.M, & Lun Lau, S. (2017). Feature selection for activity recognition from smartphone accelerometer data. Intelligent Automation and Soft Computing.

Sae-Bae, N., Ahmed, K., Isbister, K., & Memon, N. (2012). Biometric-rich gestures: a novel approach to authentication on multi-touch devices. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Austin, Texas, USA, 978-986.

Sae-Bae, N., Memon, N. (2014). Online signature verification on mobile devices. Information Forensics and Security. IEEE Transactions on, 9 (6), 933–947.

Shoaib, M., Bosch, S., Durmaz Incel, O., Scholten, H., & Havinga, P.J.M. (2014). Fusion of smartphone motion sensors for physical activity recognition. Sensors, 14, 10146–10176.

Shoaib, M., Bosch, S., Incel, O., Scholten, H., & Havinga, P.J.M. (2015). A survey of online activity recognition using mobile phones. Sensors, 15, 2059–2085.

Shoaib, M., Scholten, H., & Havinga, P.J.M. (2013). Toward physical activity recognition using smartphone sensors. In Proceedings of the 2013 IEEE 10th International Conference on Ubiquitous Intelligence & Computing and 2013 IEEE 10th International Conference on Autonomic & Trusted, Vietri sul Mere, Italy, 80–87.

Shrestha, B., Saxena, N., & Harrison, J. (2013). Wave-to-access: protecting sensitive mobile devices service via a hand waving gesture. In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in 643 Artificial Intelligence and Lecture Notes in Bioinformatics). 199–217.

Sitova, Z., Sedenka, J., Yang, Q., Peng, G., Zhou, G., Gasti, P., & Balagani, K.S. (2016). HMOG: new behavioral biometric features for continues authentication of smartphone users. IEEE Trans. Inf. Forensics Secur, 11, 877–892.

Su, Z., Tong, H., Ji, P. (2014). Activity recognition with smart phone sensors. Tsinghua Sci. Technol, 19, 235–249.

Sun, L., Zhang, D., Li, B., Guo1, B., & Li, S. (2010). Activity recognition on an accelerometer embedded mobile phone with varying positions and orientation. Ubiquitous Intelligence and Computing. UIC, Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 6406.

Tao, X., Liu, F., & Zhou, T. (2004). A novel approach to intrusion detection based on support vector data description. In Proceeding of the 30th Annual Conference of IEEE Industrial Electronics Society, 2016-2021.

Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. (2009). A detailed analysis of the KDD CUP 99 data set. Proceedings of the 2009 IEEE Symposium on Computational Intelligence in Security and Defense Applications (CISDA).

Trojahn, M., Ortmeier, F. (2013). Toward mobile authentication with keystroke dynamic on mobile phones and tables. In Advanced Information Networking and Applications Workshops (WAINA), 27th International Conference on. IEEE, 697–702.

Wankhade, K., Patka, S., and Thool, R. (2013). An overview of intrusion detection based on data mining techniques. In Proceedings of the International Conference on Communication Systems and Network Tech-nologies, Gwalior, India, 626 -629.

Whang, Z. (2018). Deep learning-based intrusion detection with adversaries. IEEE Access, Challenges and Opportunities of Big Data Against Cyber Crime, 6.

Yang, L., Guo, Y., Ding, X., Han, J., Liu, Y., Wang, C., & Hu, C. (2015). Unlocking smart phone through handwaving biometrics. IEEE Trans. Mob. Comput. 14, 1044–1055.

Zheng, N., Bai, K., Huang, H., & Wang, H. (2014). You are how you touch: user verification on smartphones via tapping behaviors. In Proceedings of the International Conference on Network Protocols, ICNP, Raleigh, NC, USA, 221–232.

Zhang, Y., Pan, G., Jia, K., Lu, M., Wang, Y., & Wu, Z. (2015). Accelerometer-based gait recognition by sparse representation of signature points with Clusters. IEEE Trans. Cybern. 45, 1864–1875.