Protection of Kids from Internet Threats: A Machine Learning Approach for Classification of Age-group Based on Typing Pattern

Soumen Roy, Utpal Roy, *Member, IAENG,* D. D. Sinha

Abstract—**The numbers of Internet users from the age group below 18 are rapidly increasing. They use the Internet for doing their homework to keep in touch with their friends. It is needless to say that rampant use of Internet has an adverse effect on the growing age children having diverse curiosity. They are vulnerable to unknown threats coming from the Internet. Many Government authorities are actively trying to protect the children from these threats. But no potential method has been applied yet. Identification of age group based on face print, hand geometry, iris texture is common. In this paper, we are interested to identify the age group (<18≤) based on analyzing the typing pattern on computer keyboard and touch screen. This method is convenient way in addition with cost effective and easy to implement with the existing systems with minor alternation. This is one approach which can be used to distinguish the minor from Internet users. The moment a user is identified to be a child or minor, the next stage of protection will be auto sensing firewall appropriate for the users. The present study aims at restricting automatically the children from Internet threats and abuse of their talent in both desktop and android environments. It has been observed that a user from age group child could be discriminated from adults by analyzing the typing style on keyboard as well as on touch screen while typing.**

Index Terms—**Keystroke Dynamics, Machine Learning, Fuzzy Rough NN (FRNN), Vaguely Quantified Rough Set (VQRS), Child Protection, libSVM**

I. Introduction

TEENEGERS spend excessive time in social network sites instead of doing their homework. They are interested to use it to keep in touch with their friends, exchange information and to make new friends. Most of them know they should not share their personal information but some of them still do. According to a survey by McAfee (a Security Technology Firm), more than 62% of children shared their personal information and 39% of their parents were unaware of it and 71% of teens hide their online behavior from their parents and 56% of parents are unaware of it [1]. Another survey in India reveals that 67% of the children under 10 had Facebook account and 82% of them received inappropriate messages [2]. It is needless to say that rampant use of Internet has an adverse effect on the growing age children having diverse curiosity. They are vulnerable to unknown threats coming from the Internet. No appropriate method is used to protect the children. No method is used to verify the age groups bellow 18 automatically.

Keystroke dynamics is a convenient measurable parameter to identify the age group bellow 18 can be used to identify the children. It is a behavioral biometric characteristic which we have learned in our life relates the issues in human identification/authentication. This is the method where people can be identified by their typing style like hand writing or voice print. Being non-invasive and cost effective nature, this method is good field of research. But the performance of keystroke dynamics is less than other popular morphological biometrics characteristics like face print, iris, finger prints recognitions due to high rate of intra class variation or high failure to enroll rate. But this technique can be used to recognize the ancillary information like age group, gender, handedness, or typing skill types or hand(s) used and emotion. This idea can meet our demand.

The science behind this technique is children's physical structure, mentality, knowledge level, experience level on a computer keyboard, neurophysiological, neuropsychological factors, reading style, keyboard position reflect the typing pattern on keyboard, which discriminate the children from adults. Basic features of keystroke patterns are the time interval between a key pressed and released, the time interval between two subsequent keys pressed and released. Now days, key pressure, finger tips size, finger placement on keyboard and keystroke sound also can be considered.

There are many timing features we can extract from the raw data of key pressed and released times recorded in millisecond unit while typing the searched text.

- KD: Time interval of key pressed and released of single key.
- RR: Interval time between two subsequent keys released.
- PP: Interval time between two subsequent keys pressed.
- RP: Interval time between one key released and next key.
- PR: Interval time between one key pressed and next key released.
- T-time: Interval time between first key pressed and

Soumen Roy is with University of Calcutta, 92 APC Road, Calcutta -700 009, India. He is also with Bagnan College, Bagnan, Howrah-711303, India (soumen.roy_2007@yahoo.co.in).

Utpal Roy was with School of Technology (SOT), Assam University, Silchar 788011, India. He is now with Visva-Bharati, Santiniketan -731235, India (roy.utpal@gmail.com).

D. D. Sinha is with University of Calcutta, 92 APC Road, Calcutta -700 009, India (devadatta.sinha@gmail.com).

last key released.

- Tri-graph-time: Interval time between one key pressed and third key released.
- Four-graph-time: Interval time between one key pressed and fourth key released.

Authentic datasets on keystroke dynamics are publicly available in the Internet or we can download on request. Among these, two of them collected the samples from children group along with adults [3, 4]. Initial dataset is designed by Bicakci et al. in the year 2014, they have collected the typing pattern samples from 51 children (age below 18) and 49 adults of 100 subjects in one session with 5 repetition for two type of text patterns (".tie5Roanl" and "Mercan Otu") through desktop computer keyboard [3]. Second dataset is collected by Abed et al. in the same year, they have collected the typing pattern samples from 11 children (age below 19) and 40 adults of 51 subjects in three sessions with time period between 3 to 30 days separating each session with 15 to 20 repetition for one text pattern ("rhu.university") through touch screen mobile device, Nokia Lumia 920 (4.5" Multi-Touch, 768×1280 (332ppi), Weight: 185g) [4].

Many statistical, distance-based and data mining or machine learning algorithms have been applied on keystroke dynamics datasets in identification/authentication techniques since 1980. These are the leading machine learning approaches were used in keystroke pattern datasets. Random Forest (RF), Fuzzy Rough NN (FRNN), Fuzzy Nearest Neighbour (FNN), SVM, Multi-Layer Perceptron (MLP), Naïve Bays (NB), Bagging, K-mean and J48 are proved to be the suitable classification methods in this domain. Many optimization techniques for feature subset selection like ACO (Ant Colony Optimization), PSO (Particle Swarm Optimization), BF (Best First) and GA (Genetic Algorithm) were used as optimization technique to select the effective features on Fuzzy Rough set. Filtered and wrapper both approaches were also applied to get the optimum feature subset so we can reach more acceptable accuracy. Some free statistical tools (Weka GUI, Orange Canvas and R Studio) are available to evaluate the classification performance and optimization techniques.

The main objective of this study is to introducing the keystroke dynamics to senegrate the minor from adult in desktop and touch screen both environments with suitable published and recognize algorithms and finding out the best suitable method which is fit for this technique and achieved the highest performance. Our objectives and contributions of the study are listed below:

- Provide a novel model to protect the children group from Internet threats based on typing pattern analysis not only through a computer keyboard also through android hand held device.
- Evaluate the performance in age group identification for recognized and published different machine learning algorithms.
- Discuss the appropriate area of application where this technique can be fit with the acceptable error rates.
- Comparative analysis of different learning methods in age group (<18≤) identification on different authenticate datasets collected through keyboard and touch screen.

This article explained the real life problem, importance and probable solutions for supporting age group identification through keystroke dynamics biometrics which has good application in real life. The proposed approach represents an interesting step forward in the field where minor or kids could be safe from looming threats from Internet. System will take the way of typing on keyboard/touch screen instead of entered texts and analyses the typing pattern and confirm the age group of the user. Keystroke dynamics and mouse dynamics are two common measurable distance-based activities to use the Internet through keyboard/touch screen. It is possible to identify the age group through keystroke dynamics with 92% of accuracy which can protect the kids or minor from looming threats coming from the Internet. This accuracy rate is impressive if enrolment phase is extremely accurate.

II. Related Works

Extraction of soft biometric or secondary biometric information (ethnicity, gender, age group, body weight, body fat…) from primary biometric characteristics (face print, finger print, typing style…) is not new. The details are listed in the Table I. The main objective of these studies is to develop a model for auto profiling the individual or to enhance the performance of biometric systems in accuracy or speed.

Many variants of authenticated datasets on keystroke dynamics with soft biometric features are created and available to the Internet for further research, we can download it or we can get it on request. Table I summarizes the details of the keystroke dynamics datasets. All these datasets we have used in our study. Here all the datasets are given by a name for identification purpose throughout this paper.

Experimental setup is the most vital to sense the behabeoural biometric characteristics like keystroke dynamics. The details of the experimental setup are listed in Table II. Validity of the datasets can be measure by this table.

Success rate to extract the soft biometric information from keystroke dynamics are listed in the Table III. Obtained results are promising.

To summarize the literature, keystroke dynamics biometric is also a good choice to automatically recognize the gender, age group <30 and ≥30 or age group ≤18 and 18+ rather than face prints, finger prints… biometric. Not only that keystroke dynamics can be also applied to recognize the emotional states, hand(s) used, and typing skill and left handed or right handed individuals. All most all the researchers used SVM with RBF as supervised machine learning method. Some of them used some optimization techniques like normalization, tuning the parameter values and feature selections to get the optimum results.

TABLE I
SUMMARY OF KEYSTROKE DYNAMICS SOFT BIOMETRIC DATASETS

Dataset Name	Study	Text Pattern	Subject	Session	Repetition	Sample Size	Feature Subsets
Dataset A	Bicakci et al. [3]	“.tie5Roanl”	100	1	5	500	KD, PP, RP
Dataset B	Bicakci et al. [3]	“Mercan Otu”	100	1	5	500	KD, PP, RP
Dataset C	Bicakci et al. [3]	“.tie5RoanlMercan Otu”	100	1	5	500	KD, PP, RP
Dataset D	El-Abed et al. [4]	“rhu.university”	51	3	15-20	951	PP, PR, RP, RR

TABLE II
EXPERIMENTAL SETUP OF THE STUDIES IN LITERATURE TO ADDRESS THE ISSUES IN SOFT BIOMETRIC TRAITS IN KEYSTROKE DYNAMICS

Study	Subject Categorization	Distribution of Examples	Experimental Setup
Bicakci et al. [3]	51 of age group <=18 and 49 of age group 18+	255 childs and 245 adults	QWERTY keyoard
El-Abed et al. [4]	11 of age group <18 and 40 of age group 19+	212 childs and 739 adults	Nokia Lumia 920

TABLE III
SUCCESS ACHIEVED BY RESEARCHERS TO RECOGNISE THE SOFT BIOMETRIC TRAITS ON KEYSTROKE DYNAMICS DATASETS

Study	Soft Biometric Traits	Categorization of Classes	Achieved Accuracy	Test Options	Classifiers	Methods
Epp et al. [5]	Emotional states	Anger and Excitation	84%	10 fold cross validation	C4.5	-
Giot et al. [6]	Gender	Male and Female	91%	5 fold cross validation	SVM	RBF
Syed-Idrus et al. [7]	Gender	Male and Female	65%-90%	50% training ratio	SVM	RBF
Syed-Idrus et al. [7]	Hand(s) used	One hand and Two hands	90%	50% training ratio	SVM	RBF
Syed-Idrus et al. [7]	Age	<30 and ≧30	65%-82%	50% training ratio	SVM	RBF
Syed-Idrus et al. [7]	Handedness	Left handed and Right handed	70%-90%	50% training ratio	SVM	RBF
Bicaki et al. [3]	Age	≦18 and 18+	88.0%-88.4%	5 fold cross valiation	SVM	RBF
Bicaki et al. [3]	Age	≦18 and 18+	88.0%-88.4%	5 fold cross valiation	SVM	RBF
Bicaki et al. [3]	Age	≦18 and 18+	88.0%-88.4%	5 fold cross valiation	SVM	RBF

As per our knowledge automatics, age group recognition based on typing pattern on a desktop computer keyboard are reported in the studies [3] and reported the accuracy only 88.0%-88.4% to recognize the child age group. No work has been yet done to recognize the child user based on typing pattern on touch screen since number of touch screen users are rapidly increasing. Our approach is to develop an efficient model to senegrate the child users from adults based on typing pattern not only based on typing pattern on a computer keyboard also on touch screen device and develop a frame work to address the problems listed in the study [3].

III. PROPOSED APPROACHES

In this section, we present our approach. We normalized each datasets in order to have input values in the range (-1, 1). Then, we used two machine learning algorithms one is SVM with Radial Basis Function (RBF) Kernel function, applied on keystroke dynamics datasets most of the times and another is FRNN with quantifiers, first time we have used to perform the classification. The computation is done using cross validation test option. To get the optimum results we tuning the parameter penalization coefficient, C and kernel parameter, γ for using SVM. In our experiment, we set C= 128 and γ=0.0625 as per guided by [8]. We also tune the parameter number of nearest neighbor, K. We set K=6 for using FRNN.

A. SVM:

Support Vector Machines (SVMs) a popular supervised machine learning methods have been introduced by Vapnik et al. [9] in 1995. Now days, SVMs have been widely studied in recognition and classification techniques to balanced datasets and have shown tremendous success in hand writing recognition to text classification. Due to the remarkable success rate SVMs are also used in keystroke dynamics not only to identify the user but also used to recognize the soft biometric information. A Support Vector based machine distinguishes imposter pattern by creating margin which separates other pattern from imposter’ which provides a learning technique for pattern recognition and regression estimation. It is common and effective for large practical problems. However, most of the time beginners get unsatisfied results due to mistake of significant steps like scaling of examples, sampling the dataset, finding the best penalty parameter and kernel parameters by tuning or grid search technique, applying feature selection mechanism etc. In this paper, we follow the guide line as per Chih-Wei Hsu et al. [8]. The combination of GA feature selection and SVM classifier is good choice in data mining research. Therefore, this combination has been applied in many studies. In our paper, we have also applied this combination but obtained results in accuracy are not improved significantly. But it reduces minimum 50% of irrelevant feature. PSO, ACO are faster, we have also applied but obtained accuracies are not significant too.

B. Fuzzy Rough Nearest Neighbour:

Fuzzy-rough nearest neighbor (FRNN) [10] classification algorithm, is an alternative to Sarkar's fuzzy rough ownership function (FRNN-O) approach [11]. But FRNN uses the nearest neighbors to construct lower and upper approximations of decision classes instead of using ownership function, and classifies test instances based on their membership to these approximations [10]. FRNN-VQRS is a new approach to FRNN. The hybridization of rough sets and fuzzy sets has focused on creating an end product that extends both contributing computing paradigms in a conventional way. But adding or deleting single element drastically change the outcome of approximations. The resulting vaguely quantified rough set (VQRS) model is closely related to Ziarko's variable precision rough set (VPRS) model [12]. Chris Corneli et al. [12] revisited the hybridization process by introducing the vague quantifiers from natural language "some" or "most" instead of "at least" or "all" into the definition of upper approximation and lower approximation which is more robust in predicting error. Here, no additional information about data is required for data analysis, only the hidden facts in data are analyzed, Rough sets are concerned with indiscernibility and fuzzy sets are concerned with vagueness.

Then Vaguely Quantified Rough Sets (VQRS) is closely related to Ziarko's variable precision rough set (VPRS) model [18]. Chris Corneli et al. revisited the hybridization process by introducing the vague quantifiers from natural language "some" or "most" instead of "at least" or "all", to decide to what extent an object belongs to the lower and upper approximation.

IV. Experimental Results

We had performed various operations with SVM and FRNN on three datasets and results are recorded in accuracy and Receiver Operating Curve (ROC) area. ROC area is a good metrics when we are working with imbalanced dataset.

TABLE IV
PERFORMANCE ACHIEVED BY PROPOSED APPROACHES

	SVM (10fold cross validation)		FRNN-VQRS (10fold cross validation)	
Datasets	Accuracy (%)	ROC	Accuracy (%)	ROC
Dataset A	91.2	0.96	Dataset A	91.2
Dataset B	89.8	0.97	Dataset B	89.8
Dataset C	90.4	0.97	Dataset C	90.4
Dataset D	82.02	0.81	Dataset D	82.02

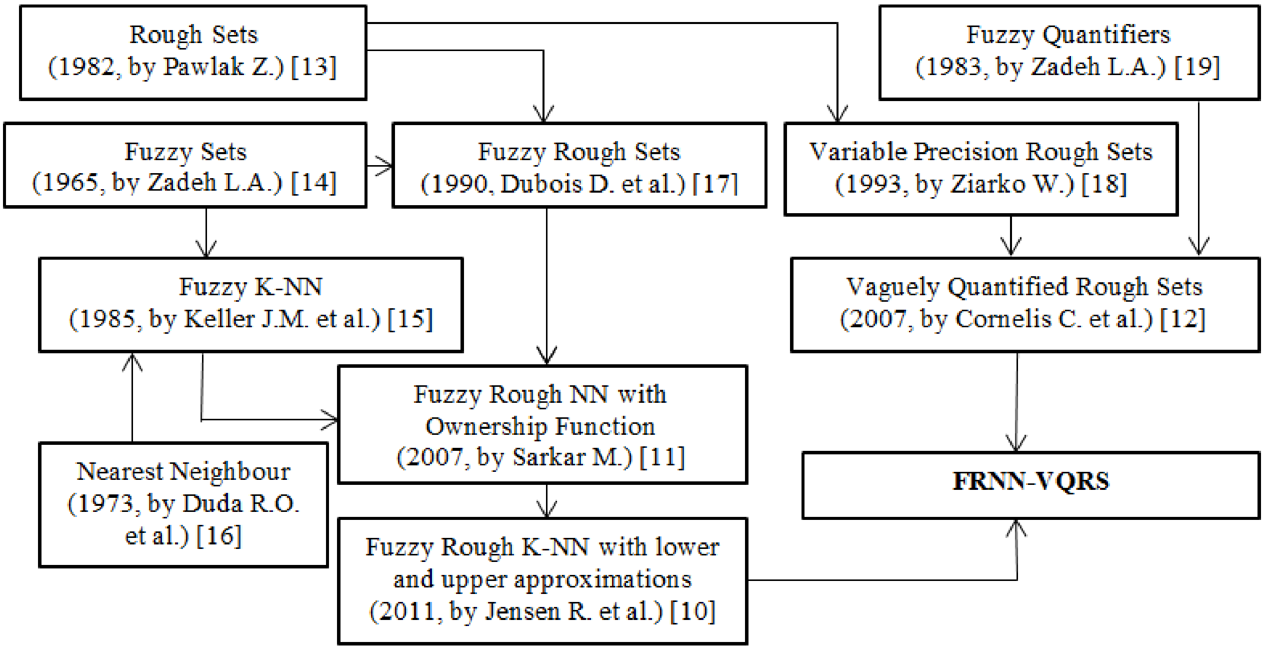
Rough Sets
(1982, by Pawlak Z.) [13]
Fuzzy Quantifiers
(1983, by Zadeh L.A.) [19]
Fuzzy Sets
(1965, by Zadeh L.A.) [14]
Fuzzy Rough Sets
(1990, Dubois D. et al.) [17]
Variable Precision Rough Sets
(1993, by Ziarko W.) [18]
Fuzzy K-NN
(1985, by Keller J.M. et al.) [15]
Vaguely Quantified Rough Sets
(2007, by Cornelis C. et al.) [12]
Fuzzy Rough NN with Ownership Function
(2007, by Sarkar M.) [11]
FRNN-VQRS
Nearest Neighbour
(1973, by Duda R.O. et al.) [16]
Fuzzy Rough K-NN with lower and upper approximations
(2011, by Jensen R. et al.) [10]

Fig. 1 - Progress of FRNN-VQRS

ISBN: 978-988-14047-8-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

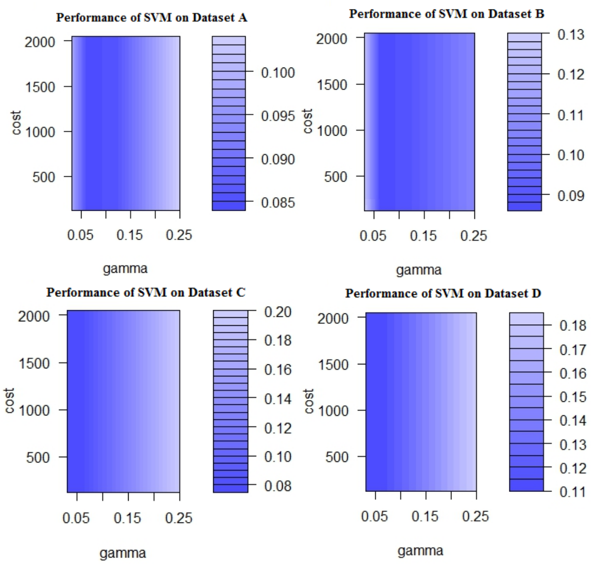

Fig. 2 - Performance of SVM on different datasets

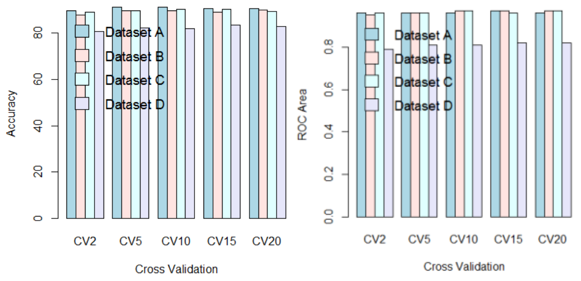

Fig. 3 - Evaluation results of SVM in different test options

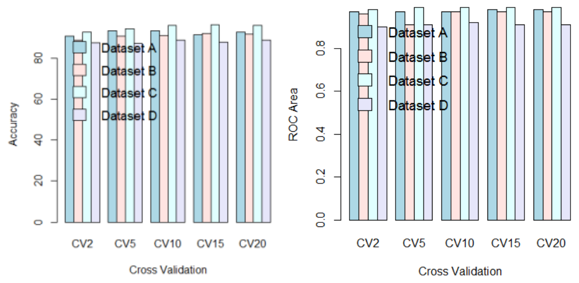

Fig. 4 - Evaluation results of FRNN-VQRS in different test options

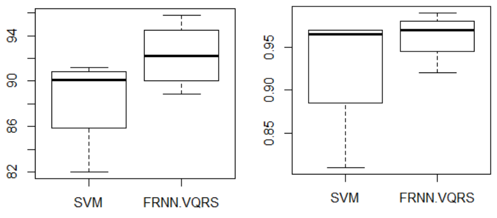

Fig. 5 - Comparison of two approaches by boxplot

Fig. 2 represents the performance of SVM against cost and gamma. The performance on Dataset D is bit low than others. In the next figure accuracy and ROC area are compared for different datasets with different cross validation (CV) test options. Fig. 4 represents the performance of FRNN-VQRS on different datasets. In the next figure, we compared the performance of the two approaches, where FRNN-VQRS is proved to be the suitable learning method in this domain.

So we can conclude that age group identification is possible in android along with the desktop environment.

V. Conclusions

This paper explained a solution for supporting age group (<18≤) identification through keystroke dynamics biometrics through a computer keyboard as well as touch screen phone which has good application in real life. Results demonstrate that proposed approach represents an interesting step forward in the field where minor or kids could be safe from looming threats from Internet. System will take the way of typing on keyboard/touch screen instead of entered texts and analyses the typing pattern and confirm the age group of the user. The moment a user is identified to be a child or minor, the next stage of protection will be auto sensing firewall appropriate for the users. It is needless to say that rampant use of Internet has an adverse effect on the growing age children having diverse curiosity. The present study aims at restricting automatically the children from Internet threats and abuse of their talent.

Many statistical, distance-based and machine learning methods have been applied on different datasets and obtained results are impressive, as per the present study, FRNN-VQRS achieved highest accuracy which is the optimum till date in desktop environment and touch screen environment. This study is a modest as well as efficient approach towards the restricted use of Internet material from the growing age children, so that they could be protected from mal use of their talent.

Keystroke dynamics and mouse dynamics are two common measurable distance-based activities to use the Internet through keyboard/touch screen. It is possible to identify the age group through keystroke dynamics which can protect the kids or minor from looming threats coming from the Internet. This accuracy rate is impressive if enrolment phase is extremely accurate. There are many factors which may affect the process and increases the failure to enroll rate. More research work has to be done and many factors have to be included like mouse dynamic, pressure which is proportional to force, depends on mass of hand weight may be the good factor in desktop environment. In android platform, key pressure, acceleration, and finger tips size may be included where advance sensing device, accelerometer are embedded in each smart phone, So this technique may get acceptable accuracy and can be used to protect the children from looming threats from Internet.

Adults hood is ascertain by attainment to 18 years of age legally. The knowledge level, IQ and ability may not always follow this suit. Exceptionally there are retarded adults as well highly proficient minors. The treatment in this paper does not discriminate the biological age. But indication is on mental age and efficiency.

Acknowledgment

Authors acknowledge all the authors listed in references.

REFERENCES

[1] McAfee, “The Digital Divide: How the Online Behavior of Teens is Getting Past Parents”, http://www.mcafee.com/in/resources/misc/digital-divide-study.pdf, June, 2012.

[2] Mugdha Variyar, “82% children on Facebook get vulgar messages”, Hindustan Times, Mumbai, http://www.hindustantimes.com/mumbai/82-children-on-facebook-get-vulgar-messages/story-0d532SUH4E2kYDN4o1ja8H.html, Feb, 2013.

[3] El-Abed M., Dafer M., El Khayat R., "RHU Keystroke : A Mobile-based Benchmark for Keystroke Dynamics Systems", 48th IEEE International Carnahan Conference on Security Technology (ICCST), Rome, Italy, 2014.

[4] Bicakci, Yuzun, “Distinguishing Child Users from Adults Using Keystroke Dynamics”, http://bil.etu.edu.tr/bicakci/dagkd/dagkd.htm, 2014.

[5] Epp, M. Lippold, and R. L. Mandryk. Identifying emotional states using keystroke dynamics. In Annual Conference on Human Factors in Computing Systems (CHI 2011), pages 715–724, May 7–12, 2011, Vancouver, BC, Canada, 2011. ACM, New York, NY.

[6] Giot R. and Rosenberger, “A New Soft Biometric Approach For Keystroke Dynamics Based on Gender Recognition”, International Journal of Information Technology and Management, Special Issue on Advances and Trends in Biometrics, Pages 1-16, 2011

[7] Idrus S. Z. S., Cherrier E, Rosenberger C. and Bours P., “Soft biometric for keystroke dynamics”, International conference on Image Analysis and Recognition, 2013, 8p.

[8] Hsu C., Chang C., Lin C. et al, A practical guide to support vector classification (2003)

[9] V. Vapnik. The Nature of Statistical Learning Theory. Springer-Verlag, New York, NY, 1995.

[10] Jensen, Richard, and Chris Cornelis. "Fuzzy-rough nearest neighbour classification." Transactions on rough sets XIII. Springer Berlin Heidelberg, 2011. 56-72.

[11] M. Sarkar, Fuzzy-rough nearest neighbor algorithms in classification, Fuzzy Sets and Systems 158 (2007) 2134 – 2152.

[12] Cornelis, Chris, Martine De Cock, and Anna Maria Radzikowska. "Vaguely quantified rough sets." Rough Sets, Fuzzy Sets, Data Mining and Granular Computing. Springer Berlin Heidelberg, 2007. 87-94

[13] Z. Pawlak, Rough Sets, Internat. J. Comput. Inform. Sci. 11, No. 5 (1982), 341-356.

[14] L.A. Zadeh, Fuzzy sets, Information and Control, vol. 8, pp. 338–353, 1965.

[15] J.M. Keller, M.R. Gray, J.A. Givens, A fuzzy K-nearest neighbor algorithm, IEEE Trans. Systems Man Cybernet. 15 (4) (1985) 580–585.

[16] R. O. Duda and P. E. Hart, “Pattern Classification and Scene Analysis,” Wiley, New York, 1973.

[17] Dubois D., and Prade H., "Rough fuzzy sets and fuzzy rough sets." International Journal of General System 17.2-3 (1990): 191-209.

[18] Ziarko W., "Variable precision rough set model." Journal of computer and system sciences 46.1 (1993): 39-59.

[19] Zadeh L. A., "A computational approach to fuzzy quantifiers in natural languages." Computers & Mathematics with applications 9.1 (1983): 149-184.